Computational Statistics for the Identification of Transcriptional Gene Regulatory Interactions

Geeven, G.

2010

document version
Publisher's PDF, also known as Version of record

Computational Statistics for the Identification of Transcriptional Gene Regulatory Interactions

Geert Geeven

About the cover:
The cover design is inspired by the title of this work. The image on the front cover is a graphical representation of a directed gene network in which the source nodes correspond to transcription factors and the target nodes represent genes. The egdes, represented by broken lines, indicate transcriptional regulatory relationships between the transcription factors and the target genes.

The publication of this thesis was financially supported by:
Vrije Universiteit Amsterdam
Thomas Stieltjes Institute for Mathematics
Netherlands Bioinformatics Centre (NBIC)

ISBN 978-90-9025640-5

Printed by Ipskamp Drukkers, Enschede

VRIJE UNIVERSITEIT

Computational Statistics for the Identification of Transcriptional Gene Regulatory Interactions

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad Doctor aan
de Vrije Universiteit Amsterdam,
op gezag van de rector magnificus
prof.dr. L.M. Bouter,
in het openbaar te verdedigen
ten overstaan van de promotiecommissie
van de faculteit der Exacte Wetenschappen
op maandag 22 november 2010 om 13.45 uur
in het auditorium van de universiteit,
De Boelelaan 1105

door

Geert Geeven

geboren te Geldrop

promotor: prof.dr. M.C.M. de Gunst
copromotor: dr. R.E. van Kesteren

Contents

1 Introduction 1
- 1.1 Biological background . . . 2
  - 1.1.1 The cell, DNA and proteins . . . 2
  - 1.1.2 Experimental data . . . 5
- 1.2 Gene networks and neuronal regeneration . . . 12
- 1.3 Outline . . . 15

2 LLM3D 17
- 2.1 Introduction . . . 18
- 2.2 Results . . . 21
- 2.3 Experimental validation of `LLM3D` predictions . . . 32
- 2.4 Discussion . . . 34
- 2.5 Methods . . . 37

3 GEMULA 49
- 3.1 Introduction . . . 50
  - 3.1.1 Regression approaches to modeling of gene expression data . . . 50
- 3.2 Methods . . . 54
  - 3.2.1 Model and notation . . . 54
  - 3.2.2 Model selection in linear models . . . 55
  - 3.2.3 Model selection criteria . . . 56
  - 3.2.4 Stepwise methods based on a selection criterion . . . 56
  - 3.2.5 Penalized least squares and the lasso . . . 57
  - 3.2.6 Random forests . . . 57
  - 3.2.7 MARS . . . 59
- 3.3 Simulation study . . . 61
  - 3.3.1 The pilot model . . . 61
  - 3.3.2 Model selection on simulated data . . . 63
  - 3.3.3 Results . . . 64
- 3.4 `GEMULA`: gene expression modeling using lasso . . . 68
- 3.5 Validation on yeast data . . . 71

3.5.1 Yeast cell cycle . . . 71
3.5.2 Yeast heat shock . . . 73
3.6 Application of GEMULA . . . 79
3.6.1 Results . . . 80
3.7 Discussion . . . 85

4 Estimation of Variable Importance **87**
4.1 Introduction . . . 88
4.2 Methods . . . 89
4.2.1 Marginal variable importance as a real-valued parameter . . . 89
4.2.2 Estimation of variable importance . . . 90
4.3 Simulation study . . . 92
4.4 Validation on yeast gene expression data . . . 96
4.5 Estimation of VIM: an application . . . 100
4.6 Discussion . . . 104

5 Conclusion **105**
5.1 The transcriptional network underlying neuronal outgrowth . . . 106

Appendix **114**
A LLM3D Supplementary Table . . . 115
B LLM3D Package Description . . . 119

Acknowledgements **129**

Samenvatting (Dutch Summary) **131**

References **133**

One

Introduction

This chapter introduces the biological concepts that are needed in the following chapters of this thesis. We give a concise and elementary discussion of the biological theory of gene expression and gene regulation for the purpose of introducing the ideas behind the models and data that are presented later on. To deal with these essentially complicated processes in a way that allows an exhibition of the fundamentals in sufficient brevity, we focus on some key concepts. For more elaborate discussions we refer to the standard textbooks on the molecular biology of the cell, such as [4, 64] and other relevant literature.

1.1 Biological background

Life on earth is extraordinarily diverse, with organisms from many different species exhibiting huge differences in size, shape and complexity of behavior. Common to all living organisms however is a remarkably complex fundamental structure which is called the cell. Cells are sometimes popularly referred to as the building blocks of life. What is important here is that every cell carries a copy of an organism's DNA, which contains hereditary information and may be viewed essentially as a blueprint or set of coded instructions for the normal development and function of the entire organism.

1.1.1 The cell, DNA and proteins

DeoxyriboNucleic Acid (DNA) consists of two long polynucleotide chains called strands. DNA molecules are of fundamental importance to life, as they contain and transfer hereditary information. Nucleotide monomers contain one of four different types of nitrogenous bases, and are commonly identified by the following single letter abbreviations: A (for adenine), C (for cytosine), T (for thymine) and G (for guanine). It is the sequence of nucleotides along the DNA strands which carries the hereditary information. The opposite ends of a strand are identified as the 5' end, marking the beginning of the strand and the 3' end marking the end. The implied direction (from 5' to 3') identifies the direction in which the encoded information is being read. With respect to a given position in a sequence, we refer to the DNA sequence toward the 5' end of the same strand as upstream and the region toward the 3' end as downstream. The two strands of the DNA run in opposite directions. The bases at opposite sides of the strands form chemical bonds. A cytosine (C) always "pairs" with a guanine (G) and an adenine (A) can only pair with a thymine (T), forming the so-called complementary base-pairs A-T and C-G. Hence, the sequence along a single strand actually conveys all coded information, and the complementarity between strands is used to copy coded information during DNA replication and transcription (see below). The complementarity between the strands is a very useful property, because two complementary single strands of DNA can "recognize" each other, which can be exploited experimentally to identify DNA sequences in a sample.

In eukaryotic species, like yeast and rat which are considered in this thesis, almost all DNA is contained in a large membrane enclosed organelle called the *nucleus*, which is absent in prokaryotic cells. Complex multi-cellular organisms, like humans and rats, are composed of trillions of different cells. Cells can be highly specialized and are organized in tissues which in turn form organs. Normal cellular functions, including growth, cell division and communication with other cells, require the synthesis of large biomolecules called *proteins*. Proteins form essential structural parts of organisms and participate in virtually every cellular process. Also known as polypeptides, proteins consist of one or more chains of amino acids. The particular interactions between amino acids in a polypeptide chain govern the proteinŠs three-dimensional structure. For most proteins, the characteristic three-dimensional structure is essential to their function. Cells can synthesize most proteins they need on demand. The specific sort and amount of proteins a cell needs is cell-type specific and condition-dependent. This makes protein synthesis a highly dynamic process. The way in which cells accomplish

this spatio-temporal *expression* of proteins is one of the most important problems studied in molecular and cellular biology. The instructions on how and when to make proteins are

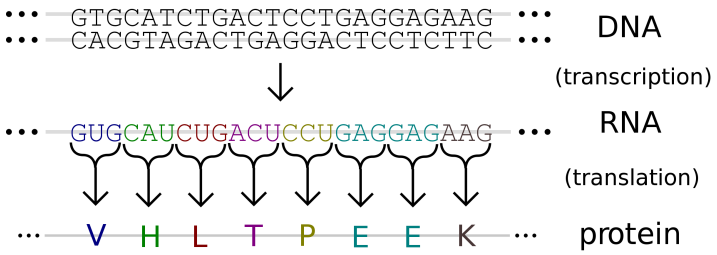

Figure 1.1: The central dogma of molecular biology and the genetic code (Picture source: Wikipedia [67]).

coded in the DNA. A fundamental theorem in molecular biology regards the encoding of information from DNA to protein, mediated by nucleic acids similar to DNA, called RNA. In brief, the *central dogma* states the following

1. DNA is *replicated* to pass information to daughter cells (through regular cell division) and eventually progeny (through the formation of gametes and sexual reproduction).

2. DNA is *transcribed* into messenger RNA (mRNA).

3. mRNA is processed and translocates from the nucleus to the cytoplasm.

4. Ribosomes in the cytoplasm interpret the mRNA code and synthesize the corresponding protein in a process called *translation*.

Figure 1.1 provides a schematic view of how genetic code contained in the DNA and the mRNA is translated into proteins. For parts of the DNA called *coding sequences*, the code defines a mapping between *codons* (three successive RNA nucleotides) and the 20 naturally occurring amino acids used in the synthesis of proteins. For instance, the codon `GUG` maps to Valine (V) and `AAG` maps to Lysine (K), see Figure 1.1.

Coding sequences represent only a small fraction of the DNA. The DNA of many organisms consists to a large extent of different sorts of repeats. Apart from coding regions and repeats, DNA sequences may have regulatory or structural functions. For large parts of the DNA the function remains to be determined. The coding regions are essential parts of structures called *genes*. Figure 1.2 depicts the typical structure of an eukaryotic gene in a higher organism. Transcription of genomic DNA leads to a primary RNA transcript that contains both coding regions (*exons*) and *intragenic* regions or *introns*. The introns are *spliced* out to form a mature mRNA which in addition to the coding region contains so-called 3' and 5' *untranslated regions* (UTRs). We return to the meaning of the promoter and enhancer genomic DNA elements, which are also indicated in Figure 1.2, later. The term *gene* is used to refer to the complete DNA sequence which is required for the production of a functional protein, and hence we speak of *transcription* and *expression* of genes in relation to the production of RNA from a

gene. It is the mechanism of regulation of transcription of genes on which we focus in this thesis.

The notion of a gene as the basic unit of heredity has been around for a long time and dates back to the revolutionary work on inheritance and genetics by Gregor Mendel (1822-1884). The discovery of DNA and the genetic code led to the first molecular characterizations of the gene, but recent discoveries driven by developments in sequencing technology have drastically changed the view on what constitutes a gene and this view is still evolving, see for instance [37] for a discussion. The term *genome* is used to denote the complete hereditary information, i.e. the complete DNA sequence, of an entire organism. An organism's genome

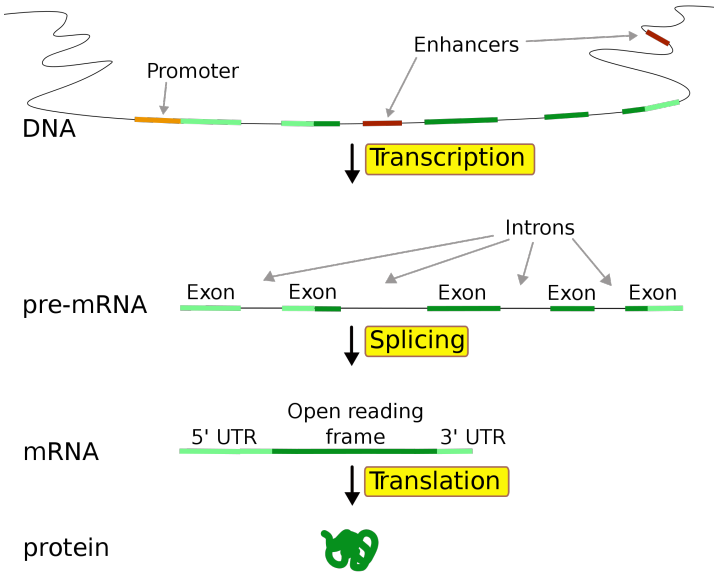

Figure 1.2: Structure of an eukaryotic gene (Picture source: Wikipedia [29]).

essentially contains coded information to make any protein any cell may need to express at any time. Several different mechanisms that cells use to regulate the expression of genes/proteins have been identified. These include transcriptional regulation, chromatin modification, DNA methylation, mRNA stability, post-translational modifications and protein degradation. Here, we only focus on regulation at the transcriptional level, i.e. on regulation of the rate at which a gene is transcribed from DNA to mRNA resulting in lower or higher observable levels of mRNA with respect to some baseline condition.

Initiation and regulation of transcription require recruitment of numerous regulatory proteins to the DNA. Important proteins are RNA polymerase II, the enzyme that synthesizes mRNA, and *transcription factors* (TFs) that bind to regulatory DNA sequences near the coding regions of the gene. The core promoter (see Figure 1.2) is a DNA sequence that contains the binding site for the basal transcription complex, which includes the RNA polymerase and additional proteins that are necessary for transcription initiation. Other DNA elements, such

as enhancers and silencers are other genomic DNA sequences containing binding sites for TFs that influence the rate of transcription of a gene. Binding of such *trans-acting* TFs to DNA may accelerate (enhancers) or decrease (silencers) the rate of transcription. Especially in

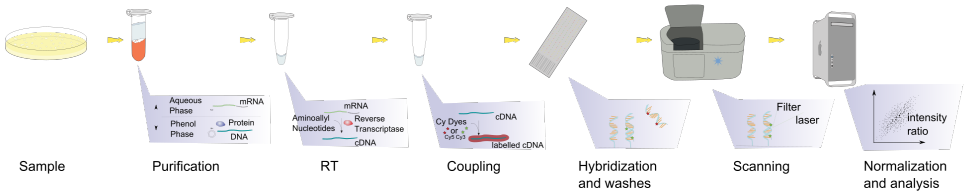

Figure 1.3: Graphical representation of the experimental steps in a typical microarray experiment (Picture source: Wikipedia [98]).

higher organisms, such as mammals, coordinated and condition-dependent action of multiple DNA binding TFs is believed to be crucial for spatio-temporal regulation on a gene-by-gene basis. Therefore, genome-wide identification of *regulatory elements* containing functional transcription factor binding sites (TFBSs) is of fundamental importance to the understanding of cellular function under normal and pathological conditions. Although it appears that the approximately 1kb (=1,000 base pairs) wide genomic region directly upstream of the transcription start site of a gene houses most functional regulatory elements, it is known that in higher organisms these elements may also be located further upstream, downstream and in intronic sequences. The size and complexity of the genome of higher eukaryotes make identification of regulatory elements more challenging for these organisms. The main goal of this thesis is to develop and apply computational and statistical methods to identify TFs that are involved in the spatio-temporal regulation of genes in higher organisms. In particular, we apply these methods to identify TFs that regulate genes that are crucial for successful regeneration of nerves after injury and neuronal outgrowth in rats.

1.1.2 Experimental data

In this thesis, we present several approaches to modeling experimental data aimed at elucidating mechanisms of transcriptional regulation. The primary sources of experimental data we use in our models are

1. Gene expression measured with DNA microarrays.
2. Predicted binding sites/binding affinities of DNA binding TFs in regulatory DNA sequences surrounding coding sequences of genes.
3. Functional genomics data, such as functional gene annotations as provided by the Gene Ontology (GO) consortium [21].

In the remainder of this section we briefly describe the nature of these three main different data types. Apart from these, any functional biological data we may obtain for a sufficiently large population of genes that is relevant to the (transcriptional) regulation of the genes may be used in addition. We will come back to this issue in Chapter 3.

1.1.2.1 DNA microarrays and gene expression

A cell's transcriptome is defined as the total pool of mRNAs present in the cell and reflecting the set of active genes in the cell. DNA microarray technology allows the quantification of a cell's transcriptome and, in particular, changes in the transcriptome as a function of intra- and extracellular conditions. Two-color microarrays, for example, allow the quantification of relative amounts of mRNA in two different biological samples. They can be used, for instance, to compare the expression of genes in "normal" or "healthy" cells with expression in "diseased" cells, or cells that have undergone some sort of treatment. Proteins encoded by genes that show significant changes in expression may be potential targets for novel drugs or may function as markers, e.g. to classify and diagnose different types of tumors. There are many microarray platforms around nowadays that differ with respect to design, accuracy, efficiency and cost. Moreover, studies employing microarray technology may differ substantially in experimental design and protocols used. A typical microarray experiment involves the following steps

1. Extraction of mRNA from biological samples.
2. Preprocessing and labeling of mRNA samples.
3. Hybridization of labeled samples to a microarray chip.
4. Quantification of (relative) amounts of label hybridized to DNA on the chip by laser scanning.
5. Processing of the resulting raw measurements to produce data ready for higher level analysis.

Figure 1.3 illustrates these steps. The above mentioned procedure is a highly complex technological process where in different steps both biological and technological variation influence the outcome. As a result, reproducibility is a serious issue in microarray experiments and the importance of careful design and execution of the experiments and processing of the results can hardly be overstated.

1.1.2.2 DNA binding transcription factors: prediction of binding sites and binding affinities

TFs contain DNA binding domains which bind DNA in a sequence specific manner. Figure 1.4 shows a cartoon representation of an experimentally derived three dimensional structure of a TF-DNA complex. TF binding sites on the DNA can be determined experimentally using different techniques, including chromatin immunoprecipitation (ChIP), DNA footprinting and Systematic Evolution of Ligands by Exponential Enrichment (SELEX). Binding sites for TFs are typically 6 to 20 base pairs (bp) in length, highly degenerate, and there is significant overlap in sequences bound by structurally related but different TFs. This makes genome-wide computational prediction of binding sites and target genes of TFs challenging. *In silico* prediction of new potential binding sites of a TF in DNA sequences usually involves the following steps.

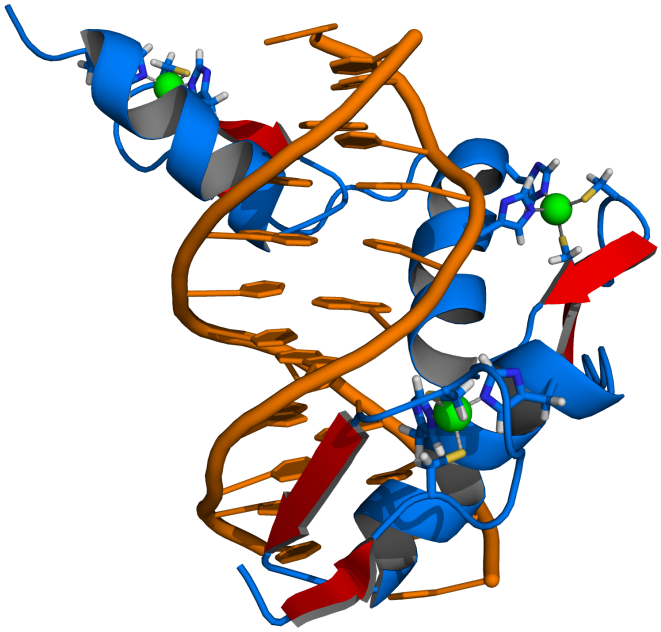

Figure 1.4: Cartoon representation of the TF `EGR1` bound to DNA (Picture source: Wikipedia [97]).

1. Experimental characterization of sequences of DNA binding sites of the TF.
2. Derivation of some probabilistic or biophysical model of the DNA binding affinity of the TF for every possible DNA sequence.
3. Scanning of the genome for DNA sequences to which the TF is *likely* to bind based on the derived model.

Several databases, such as TRANSFAC [68], Jaspar [15] and YEASTRACT [104], contain experimentally derived binding site models for a large number of TFs. As an example, we consider the DNA binding protein *cAMP response element binding* (CREB). CREB is a crucial regulator of many cellular processes. Figure 1.5 lists a subset of CREB binding sites present in TRANSFAC along with a so-called *Position Frequency Matrix* (PFM). Suppose we observe 29 DNA binding site sequences, all of which have a length of 8 bp. For each position $j \in \{1,\dots,8\}$ in each binding site k, for $k \in \{1,\dots,29\}$, let x_{jk} be the nucleotide at position j in binding site k. Hence, $x_{jk} \in \{A,C,G,T\}, \forall j,k$. The PFM contains counts of observed nucleotides at positions j of the binding site and, for $i \in \{A,C,G,T\}$, the y_{ij} entry of the PFM in Figure 1.5 is

$$y_{ij} = \sum_{k=1}^{29} |\{x_{jk}\} \cap \{i\}|.$$

This information can be used to construct a probabilistic model for the binding site. The DNA sequences to which a TF binds are sometimes referred to as the DNA *motif* of the TF. The purpose of the model for the binding site is to discover potential new binding sites in

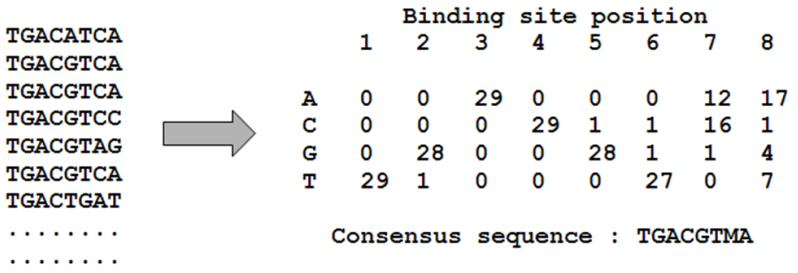

Figure 1.5: CREB binding sites, PFM and consensus sequence (data from TRANSFAC [68]).

a genomic DNA sequence. Let $\mathbf{S} = (S_1, \ldots, S_L)$, where $S_l \in \{A, C, G, T\}$, for $l = 1, \ldots, L$, be a genomic DNA sequence of length L. Typically $L >> w$, where w is the width of the DNA motif. The model that is used to find binding sites is composed of two submodels, one for observing nucleotides in so-called *background sequence* and one for observing nucleotides in the binding site. Let $b_j(i)$ denote the probability of observing nucleotide i at position j in a binding site and let $q(i)$ denote the probability of observing nucleotide i in a stretch of background sequence. For a subsequence $\mathbf{S}_l = (S_l, \ldots, S_{l+w-1})$ of length w in $\mathbf{S}$ starting at position l, a binding site similarity score can be computed as

$$z_l = \sum_{j=1}^{w} \log\left(\frac{b_j(S_{l+j-1})}{q(S_{l+j-1})}\right).$$

The score can be computed for all $l = 1, \ldots, L - w + 1$ and compared to a suitably chosen threshold to predict whether $\mathbf{S}$ contains any possible binding sites. For this purpose, the PFM is converted into a *Position Specific Scoring Matrix*(PSSM), containing the $\log(\frac{b_j(i)}{q(i)})$ as entries, for $i \in \{A, C, G, T\}$. Figure 1.6 contains an example of an experimentally determined PSSM for the CREB binding site. The exact details regarding the design and estimation of PSSM models differ slightly between commonly used implementations, see for instance [100, 40, 80] for more details.

The DNA motif of a TF is often represented as either a consensus sequence or a sequence logo. The consensus sequence shows which nucleotide(s) is (are) most abundant in the binding site at each position. Figure 1.5 shows the consensus sequence for the CREB binding site. The letter `M` is used to denote nucleotide `A` *or* `C`. A sequence logo is a graphical representation of the motif in which the height of the letters at each position is proportional to the information content at that position. The information content at a position reflects the

	Binding site position							
	1	2	3	4	5	6	7	8
A	0.00	0.00	2900.00	0.00	0.00	0.00	499.41	417.87
C	0.00	0.00	0.00	2900.00	89.18	78.45	665.88	24.58
G	0.00	2497.04	0.00	0.00	2497.04	78.45	41.62	98.32
T	2900.00	89.18	0.00	0.00	0.00	2118.13	0.00	172.07

Position-specific scoring matrix

Figure 1.6: PSSM and sequence logo for the CREB binding site (data from TRANSFAC [68]).

non-degeneracy in the motif. A logo of the CREB binding site is shown in Figure 1.6. For more details on consensus sequences and logos we refer to Schneider *et al.* [89].

To a certain extent, TFs are organism specific. That is, the genomes of plants, bacteria, fungi and vertebrates generally encode different sets of TFs. However, the structural properties of TFs are reasonably well conserved between closely related species. For instance, most TFs in human, mouse and rat are very similar. For these species, the collection of vertebrate PSSMs from TRANSFAC [68] represent a fairly complete set of currently known motifs of DNA binding TFs. Given a collection of PSSMs and a set of genomic DNA sequences that represent regulatory regions of known genes, binding sites can be predicted *in silico*. A schematic overview of an approach to genome-wide computational prediction of TFBS is presented in Figure 1.7. The resulting predictions can be used in models that relate the presence of binding sites of TFs in the promoters of genes to the observed expression of those genes (see Chapter 2).

From a biophysical point of view, modeling binding of a transcription factor to the DNA as a discrete event is a simplification. In an attempt to exploit experimental and theoretical knowledge regarding protein-DNA binding, Roider *et al.* [85] developed a biophysical model, called `TRAP`, for prediction of transcription factor binding affinity. `TRAP` avoids the artificial separation between binding sites and non-binding sites. Instead, `TRAP` computes the binding probability of a given TF to each possible site in the sequence. These binding probabilities are summed over all positions in a sequence to give an estimate of the total binding affinity of the TF for a given promoter. The binding affinity derived from this model is a continuous measure and can be used to quantify binding of TFs to TFBSs more accurately than discrete TFBS descriptions. The affinity predictions can be calibrated to reproduce experimental binding data when available, but also allow for prediction solely based on matrix descriptions of a binding site, as given in the previous section. We expect `TRAP` affinities, being continuous measures, to be useful particularly as predictors in regression models, which we present in Chapters 3 and 4.

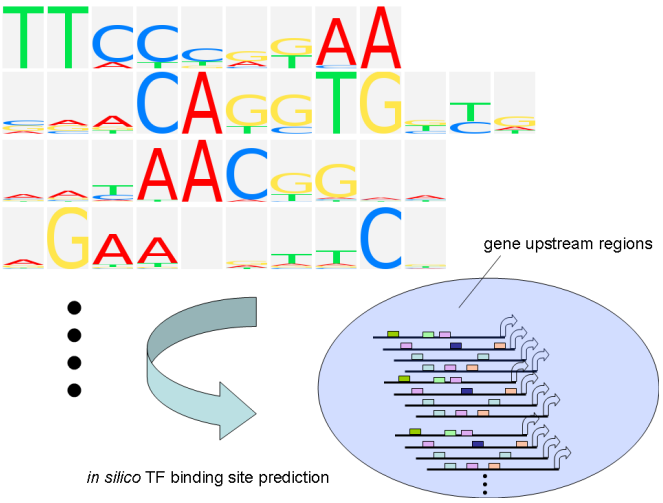

Figure 1.7: Predicting TFBSs *in silico* using PSSMs.

1.1.2.3 Gene Ontology

The Gene Ontology (GO) project is a major collaborative initiative that provides a set of structured, controlled vocabularies for general use in annotating genes, gene products and sequences [7]. The project is maintained by the Gene Ontology Consortium whose main aim is to produce a dynamic vocabulary that can be applied to all eukaryotes even as knowledge of gene and protein roles in cells is accumulating and changing. The GO project has developed three structured controlled ontologies that describe gene products in terms of their associated biological processes, cellular components and molecular functions.

1. Cellular Component (CC); GO terms describing the different parts of a cell or its extracellular environment to which gene products are localized.

2. Molecular Function (MF); GO terms describing the elemental functional activities of a gene product at the molecular level.

3. Biological Process (BP); GO terms describing operations or sets of molecular events that gene products take part in. These events have a defined beginning and end and are pertinent to the functioning of integrated living units: cells, tissues, organs, and organisms.

Each ontology is structured as a directed acyclic graph (DAG), in which child nodes and parent nodes are related through relationships such as "is-a" and "part-of". For instance, the GO term `metabolic process` is related to its parent GO term `biological process` by an "is-a" relationship. A notable feature of the GO DAG is that it creates a hierarchy of GO terms where high level terms close to the root node represent general terms that may describe many genes, whereas low level terms near the end nodes of the graph are very

specific. This creates redundancy in the GO terms as the sets of genes annotated to closely related terms overlap significantly. We use the GO BP annotations of genes in Chapter 2 to group together genes that are involved in the same biological process.

1.2 Gene networks and neuronal regeneration

The emergence of high-throughput technologies in the -omics era has enabled biologist to characterize biological systems experimentally in great detail and this has created a wealth of data on a genome-wide scale. While ever more data is becoming available, a subject of increasing interest to biologists is the study of biological networks (Barabasi and Oltvai [10], Hayete *et al.* [43], Jothi *et al.* [52]), which describe different kinds of interactions between genes and proteins. Modeling interactions between genes and proteins is believed to be crucial for the understanding of biological systems. Here, we distinguish between two different kinds of gene networks.

1. Gene coexpression networks (GCN)s. In a GCN, nodes represent genes and an undirected edge between two nodes indicates that the two genes are coexpressed in some biological condition of interest. As coexpressed genes are often functionally related, GCNs can be used to functionally characterize sets of genes involved in a biological process or to predict novel functional roles of previously uncharacterized genes (Stuart *et al.* [101], Lee *et al.* [58]).

2. Gene regulatory networks (GRN)s. A GRN is a directed network, with edges between source nodes and target nodes. The source nodes are TFs and the target nodes are transcriptional targets. An edge between a TF and a target gene may represent either the binding of the TF to the target gene's promoter, a direct influence of the TF on the rate of transcription of the target or (preferably) both (Lee *et al.* [59], Segal *et al.* [91], Harbison *et al.* [41]).

Biologically, both types of networks are of interest and in fact the two types of networks can be integrated and no strict distinction between GCNs and GRNs is necessary. However, we make the distinction here, because the type of experimental data available and the computational method used for inference of the network, determine what kind of interactions can be modeled. Over the past decade, several different computational statistical methods have emerged for the inference of gene networks from data (see Hecker *et al.* [44] and Lee and Tzou [60] for two recent reviews). Given our interest in modeling the gene network that underlies successful neuronal outgrowth, initially our focus was on learning network structures from gene expression data using graphical models. Graphical models provide a way of modeling conditional (in)dependencies between all genes jointly and they have been used previously to model large networks of interacting genes (Friedman *et al.* [31], Hartemink [42], Wehrli *et al.* [114], Schäfer and Strimmer [86]). Within the class of graphical models, two special types of models are popular for the application of gene network inference. These are graphical Gaussian models (GGMs) (see Lauritzen [57], Schäfer and Strimmer [86, 87], Wehrli *et al.* [114]) and Bayesian networks (BNs) (see Jordan [51], Friedman *et al.* [31], Hartemink [42]). Graphical Gaussian models are undirected probabilistic models in which the joint distribution of the collection of random variables in the model is a multivariate Gaussian distribution (see Whittaker [116] for a basic introduction). GGMs can be represented by graphs in which an edge between two nodes represents a dependence, i.e. two unconnected nodes are assumed conditionally independent, where the conditioning set consists of all remaining nodes in the network. From GGM theory it is well known

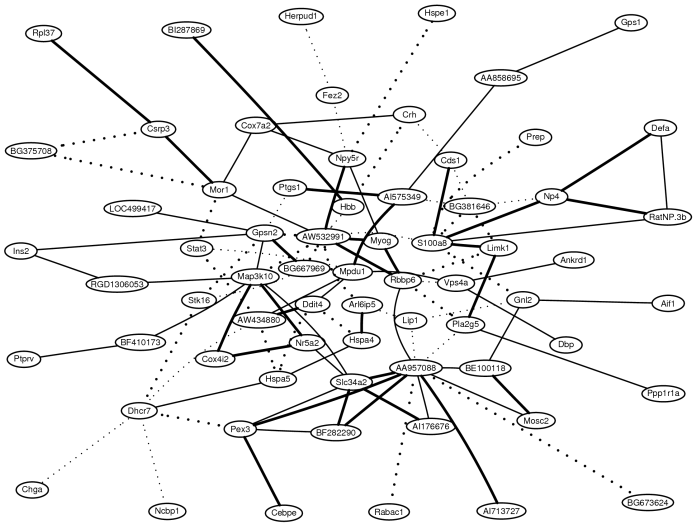

Figure 1.8: A GGM model inferred from 100 potential regeneration associated genes in rat DRG neurons following DR crush. We analyze gene expression from this *in vivo* regeneration model in Chapter 2.

that such direct dependences correspond to the non-zero entries in the partial correlation matrix, which is related to the inverse of the covariance matrix of the joint random vector of network nodes. Hence, inference of GGMs requires reliable estimation of this covariance matrix. In typical applications, the number of nodes p in the network exceeds the sample size n available for estimation. Schäfer and Strimmer [87] proposed the use of shrinkage estimators for the stable estimation of the covariance matrix and developed `GeneNet`, an `R` package for the inference of large-scale gene association networks from gene expression data.

A BN is a probabilistic graphical model based on a directed acyclic graph (DAG) that represents conditional dependencies between random variables, which are the nodes in the DAG. The joint distribution of variables is determined by the DAG, a family F of (conditional) probability distributions and parameters θ that correspond to the distributions in F. The joint distribution of the nodes factorizes according to the DAG in a special way and each variable is conditionally independent of its nondescendants, given its direct parents. A notable feature of BNs is that prior knowledge can be incorporated in a prior distribution over all possible DAGs. Structure inference is usually based upon a Bayesian score, which is determined by both the prior and the observed data. Werhli *et al.* [114] published a comparative evaluation of gene network inference using GGMs and BNs on both real experimental data and simulated data. Werhli *et al.* conclude in [114] that with passively observed gene expression data, directions can not be distinguished. Given the high computational demand of BN inference and the marginal differences in performance on benchmark datasets, Werhli *et al.* recommend

GGMs over BNs for inference of GCNs, unless the gene expression data was obtained after perturbations in the system.

The biological application we are mainly interested in is the gene regulatory network underlying neuronal regeneration. In this thesis, we analyze data from different *biological* models of this process. Dorsal root ganglion (DRG) neurons display robust and successful regeneration following lesion of their peripheral neurite, whereas outgrowth of lesioned central neurites is weak and does not lead to functional recovery. The DRG is therefore an excellent *in vivo* model system for regeneration since it allows one to study both successful and unsuccessful regeneration and to characterize their differences. The F11 cell line is a fusion product of mouse neuroblastoma cells with embryonic rat DRG neurons. Upon stimulation with Forskolin, a chemical agent which raises intracellular levels of cAMP, F11 cells acquire a neuronal phenotype which results in the outgrowth of neurites. F11 cells are easy to culture and transfect and provide a good *in vitro* model for the transcriptional regulation of DRG regeneration *in vivo* (MacGillavry *et al.* [65]). However, it is a strongly reduced model in which both the lesion stimulus and the DRG cellular environment are lacking. The intrinsic potential of neurons to regrow damaged nerve fibers after an injury depends in part on their ability to initiate a growth promoting gene expression program. Coordinated expression of regeneration associated genes is believed to be governed by a temporally dynamic network that contains interactions between TFs and target genes. Through analysis of gene expression and DNA sequence data of regeneration associated genes from both *in vivo* and *in vitro* biological models of regeneration, we aim to identify TF-target-gene regulatory interactions that are crucial for robust and successful neurite outgrowth.

Initially, we attempted to infer a GRN of genes encoding TFs and putative target genes associated to robust and successful outgrowth of neurites using GGMs and BNs. A typical example of a resulting network is depicted in Figure 1.8. The nodes in this network are the 100 most strongly down-regulated genes in rat DRG neurons in response to a lesion of the central neurites, constituting putative regeneration-inhibiting genes, complemented with several TFs which have been identified as potential regeneration associated TFs by Stam *et al.* [99]. Regulatory relationships can be established by looking at genes that are coexpressed with TFs of interest. However, we found that TFs that are likely to be important for coordination of regeneration associated gene expression are not always regulated at the transcriptional level and that the inferred networks often contain edges between genes that do not encode TF proteins. Moreover, the networks that we inferred in this way are based exclusively on gene expression data and do not take into account the effect of binding of TFs to TFBSs on gene expression of targets. Hence, preliminary results of analysis using graphical models did not provide us with enough TF-target-gene relationships that were promising for experimental validation. Therefore, in this thesis we decided to focus on the development and improvement of computational statistical methods that can be used to infer TF-target-gene relationships directly and model the effect of binding of TFs to TFBSs on variation in gene expression.

1.3 Outline

The remainder of this thesis is organized as follows. In Chapter 2 we present a novel method (LLM3D) that uses log-linear modeling of three-dimensional contingency tables to predict transcriptional regulators of functionally homogeneous and condition-specific sets of target genes from genome-wide expression data. LLM3D simultaneously uses gene expression, gene ontology (GO) annotation and computationally predicted binding sites (TFBS) in a combined statistical analysis based on log-linear models, and is aimed at finding TFBS-GO pairs that are significantly associated with a gene expression response of interest.

In Chapter 3 we study regression approaches to modeling of gene expression data. Regression models can be used to quantify the amount of variation in gene expression that can be explained by biologically relevant covariates such as predicted binding affinity of TFs. These models can also be used to identify combinatorial regulation by modeling the joint effect of multiple TFs on gene expression through interactions. We propose GEMULA, a strategy based on linear models that is fast, considers a wide range of biologically plausible models and selects parsimonious and interpretable models from experimental data.

In Chapter 4 we develop a statistical approach to the estimation of individual marginal effects of predictors on gene expression when a model in which multiple predictors are present is given. The goal is to obtain a sensible ranking of the predictors that reflects the relative contribution of predictors in explaining variation in gene expression. We use a statistical framework for variable importance estimation to define the marginal importance of predictors as a parameter and show that this parameter has an intuitive and interesting biological interpretation.

We conclude in Chapter 5 by applying the techniques we developed in Chapter 2, 3 and 4 to experimental data from an *in vitro* biological model of neuronal regeneration and integrate the results into a temporally dynamic transcriptional network underlying neuronal outgrowth.

Two

LLM3D

We developed a new method that uses log-linear modeling of three-dimensional contingency tables (LLM3D), to predict transcriptional regulators of functionally homogeneous and condition-specific sets of target genes from genome-wide expression data. LLM3D simultaneously uses gene expression data, gene ontology (GO) annotation and computationally predicted transcription factor binding sites (TFBSs) in a combined statistical analysis based on log-linear models, and is aimed at finding TFBS-GO pairs that are significantly associated with a gene expression response of interest. LLM3D offers a methodological improvement over existing enrichment-based methods because it achieves significantly higher statistical power to detect biologically relevant gene expression-TFBS-GO relationships. Furthermore, LLM3D is generally applicable and can be readily adapted to any context or organism. Using published data on yeast and human gene expression and on transcription factor binding, we were able to validate LLM3D and demonstrate a significant improvement in performance compared with existing methods. We further showcase LLM3D performance by identifying and experimentally validating novel gene regulatory interactions involved in the regenerative growth of injured mammalian neurons.

2.1 Introduction

Condition-specific and time-dependent transcriptional regulatory networks underlie the coordinated expression of genes involved in all biological processes. Insight into these networks is crucial for the understanding of biological systems under normal and pathological conditions. An important step in studying gene regulatory networks is the prediction of functional transcription factor binding sites (TFBSs) within gene regulatory sequences. Recently, advanced methods have been developed to predict TFBSs *in silico* (Hannenhalli [40], Wasserman and Sandelin [113]). Public databases containing large collections of experimentally validated binding sites can be used to derive probabilistic models of TFBSs and software algorithms can subsequently be employed to scan potential gene regulatory sequences for the prediction of new sites. However, whereas in simple model organisms such as yeast, gene regulatory sequences are well defined and relatively small in size, mammalian gene regulatory sequences are large and can be located up to several thousands of base pairs away from transcription start sites. Consequently, mammalian TFBS predictions are usually less accurate and more likely to contain false-positives. False-positive predictions can be reduced by improving the predictive value of the binding sites. This can for instance be achieved by considering TF binding affinity (Roider *et al.* [85], Ward and Bussemaker [111]), TF cooperativity at *cis*-regulatory modules (Warner *et al.* [112], Zinzen *et al.* [122]) or evolutionary conservation of binding sites across species (Wasserman and Sandelin [113], Xie *et al.* [120]). Here, we describe a novel computational method that improves the use of TFBS descriptions through gene set analysis based on enrichment.

A commonly used method to reduce false-positive TFBS predictions at the computational level involves the identification of TFBSs that are enriched in subsets of related genes compared to a control (background) set of genes. Over the past decade, many different gene set enrichment tools have been developed (Huang da *et al.* [50], Nam and Kim [72]). Under the assumption that common transcriptional regulation underlies the co-expression of functionally coherent sets of genes, co-regulation and co-functionality are often used as criteria to define gene sets of interest. In order to study enrichment of both TFBSs and gene function in co-expressed genes, two different computational approaches can be used. The first approach, which is referred to as singular enrichment analysis (SEA) [50], allows separate quantification of gene ontology (GO) term and TFBS enrichment in sets of co-expressed genes (Figure 2.1((**a**))). For any given gene set, enrichment is tested for each GO term and TFBS independently, and in case of multiple input gene sets, the procedure is simply repeated for each set. Examples of methods that employ a SEA strategy include `DAVID` (Huang da *et al.* [49]) and `g:Profiler` (Reimand *et al.* [82]). SEA analysis typically returns separate lists of enriched features (such as GO terms and TFBSs), but SEA based methods are not designed to predict transcriptional targets using gene expression, TFBS and GO data simultaneously. Therefore, we do not consider SEA for comparative data analysis in the present study.

The second approach, which we will refer to as multi gene set by intersection (MGSI), analyzes enrichment in gene sets of co-occuring biological features such as TFBSs and GO annotations. MGSI predefines multiple sets of co-expressed genes sharing the same GO, and subsequently tests each set for TFBS enrichment (Figure 2.1((**b**)). An example of a method

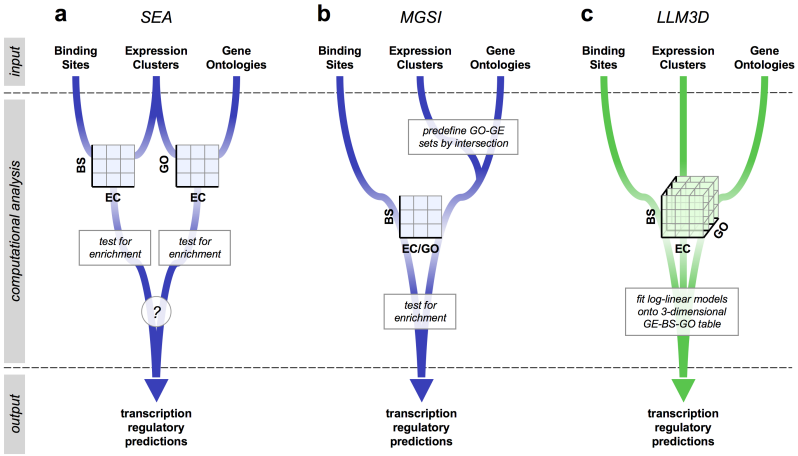

Figure 2.1: Comparison of LLM3D with other gene set enrichment analysis approaches. (**a**) In singular enrichment analysis (SEA), gene expression clusters (EC) are independently tested for enrichment of binding sites (BS) and gene ontology (GO) terms using two two-dimensional contingency tables. It is not clear how meaningful relationships between the two should be inferred. ((**b**) In multi gene set by intersection (MGSI), multiple gene sets are predefined based on intersecting sets of co-expressed genes with sets of genes sharing GO terms. MGSI considers all three variables in a single two-dimensional contingency table in which gene expression and GO data are collapsed into a single combined variable. ((**c**) In LLM3D, gene expression, binding site data and GO annotations are used as separate input variables in a single three-dimensional contingency table. To this table, log-linear models are fitted in order to test the joint dependence of all three variables simultaneously.

that employs MGSI is GeneCodis (Carmona-Saez *et al.* [74]), and the MGSI method we use here is comparable to testing significance of pairwise co-occurences of TFBSs and GO terms using GeneCodis. Although the pre-selection of gene sets that share a common function does improve the subsequent prediction of functional TFBSs, MGSI collapses gene expression and GO annotation into a single combined variable. As a result, important information about the joint dependence of all three variables (i.e., gene expression, GO association, and TFBS presence) is lost, and the statistical power to detect biologically meaningful associations is compromised.

To overcome these limitations, we developed a novel method that uses log-linear modeling of three-dimensional contingency tables (LLM3D), to identify experiment-specific associations between gene expression, TFBS presence and gene function (Figure 2.1((**c**)). We show that LLM3D provides a significant improvement over existing methods. We validate our method using published yeast and human genomewide gene expression and transcription factor binding data. We demonstrate that the gene regulatory predictions made by LLM3D are more accurate and have significantly higher predictive value compared to existing methods. Finally, we showcase LLM3D by performing and analyzing a genome-wide expression profiling study in an animal model for the functional regeneration of injured neurons. *Post hoc* experimental validation shows that in this case LLM3D is able to identify functional gene regulatory

interactions that remain undetected using existing methodology.

2.2 Results

LLM3D: a methodological and statistical improvement in gene set enrichment analysis

Input to the main statistical analysis in LLM3D is a defined set of gene expression clusters, TFBSs and GO terms. For each TFBS-GO pair of interest, LLM3D crossclassifies all genes according to GO annotation, TFBS presence, and gene expression to obtain a three dimensional contingency table. The main statistical analysis of LLM3D consists of finding a good model to describe the observed counts in this table. The most basic model, i.e., the model that assumes that gene expression, GO annotation and TFBS presence are mutually independent, is referred to as the null model. The underlying hypothesis of mutual independence is tested using a likelihood ratio test statistic (Christensen [19]). When this hypothesis is rejected, LLM3D considers eight alternative models that differ in the dependence relationships they imply between gene expression, GO annotation, TFBS presence, and selects the best model using Akaike's information criterion (AIC)(Akaike [2]). The enrichment of TFBSs in functionally homogenous and co-expressed genes implied by the selected model is then used to predict functional gene regulatory interactions. Because we consider all three variables jointly, we expect LLM3D to perform better in comparison with existing enrichment based methods. A detailed description of LLM3D is provided in Section 2.5. LLM3D is available as an R package (see Appendix B).

LLM3D identifies functional gene regulatory interactions in yeast more accurately than existing methods

To compare the predictive performance of LLM3D with that of MGSI, we applied both methods to a publicly available high-quality genome-wide time-course gene expression data set obtained in yeast. Transcriptional regulation in yeast has been studied extensively, both computationally and experimentally, and for many TFBSs, high quality PSSMs as well as physical TF-TFBS interaction data under various experimental conditions are available. We used LLM3D and MGSI to predict target genes of TFs that control the yeast metabolic cycle (Tu *et al.* [108]). It is estimated that approximately half of all yeast genes show periodic expression during the metabolic cycle. These genes can be divided into three large expression clusters of tightly co-regulated genes: an oxidative (Ox) cluster, a reductive building (Rb) cluster, and a reductive charging (Rc) cluster, and many TFs have been implicated in the periodic expression of these genes (Tu *et al.* [108]). A plot of the average gene expression of genes in each of these three clusters as a function of time during three successive metabolic cycles is shown in Figure 2.2.

We reanalyzed the yeast Ox, Rb and Rc clusters using LLM3D and MGSI, and we subsequently used two independent and complementary data sets of *in vivo* yeast TF-target gene interactions to validate our predictions. The refined regulatory map published by MacIsaac *et al.* [66] (named MRM hereafter) contains 7,840 experimentally validated interactions between 3,107 different yeast promoters and 118 yeast TFs. All interactions reported in MRM are based on chromatin immunoprecipitation (ChIP) data at a significance level of 0.001.

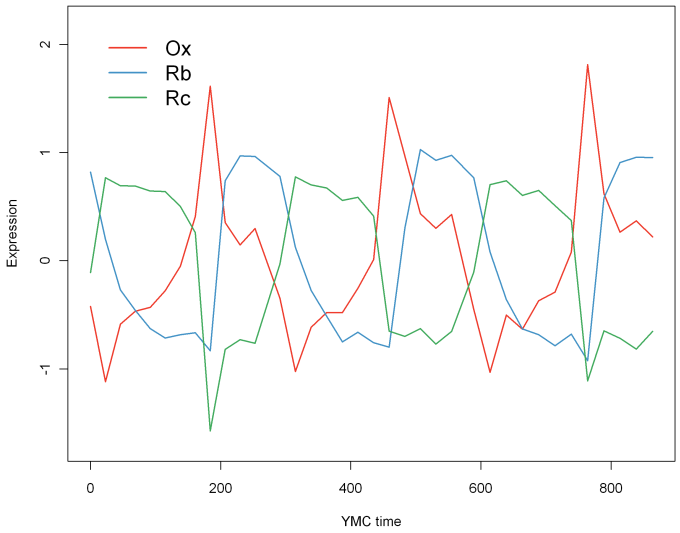

Figure 2.2: Yeast metabolic cycle gene expression.

YEASTRACT on the other hand is a curated repository of regulatory associations between TFs and target genes based on >1,000 bibliographic references (Teixera *et al.* [104]). For the 110 TFs in YEASTRACT for which PSSMs are available, the database reports 29,263 regulatory interactions with 5,913 different yeast genes. We used either YEASTRACT or MRM as a repository of true TF-target gene interactions to compare the predictive performance of `LLM3D` and MGSI on the yeast metabolic cycle gene expression data set.

Under the assumption that YEASTRACT and MRM indeed contain true TF-target gene interactions, the regulatory interactions inference task can be viewed as a binary classification problem, and predictive performance can be measured using receiver operating characteristic (ROC) curves. Regulatory predictions are classified as either true positives or false positives depending on whether predictions are confirmed in MRM or YEASTRACT. Conversely, negative predictions are either classified as true negatives or false negatives. In ROC curves the true positive rate is plotted against the false positive rate, and the area under curve (AUC) is used as a summary statistic to compare different curves. A higher AUC indicates better predictive performance.

The essential difference between the `LLM3D` approach and MGSI is depicted in Figure 2.5. In the given example, ACE2 binding sites are not detected in association with `mitosis` genes in the three yeast metabolic cycle expression clusters separately using MGSI, whereas `LLM3D` reveals a significant association of `ACE2` binding sites with `mitosis` genes considering all expression clusters simultaneously. To be able to directly compare MGSI and `LLM3D` predictive performance, we only considered the top 20 TFs for which both methods predicted

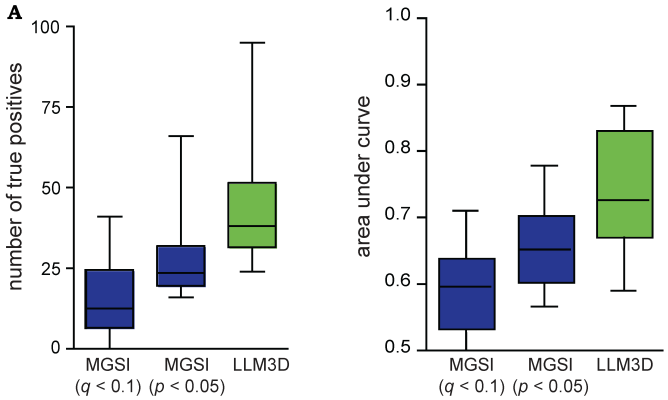

Figure 2.3: Predictive performance of `LLM3D` compared with existing methods. (**A**) Number of true positive target gene predictions for top-20 TFs identified by both `LLM3D` and MGSI in the yeast metabolic cycle expression clusters. `LLM3D` identifies more true targets than MGSI, even when the stringency of the latter is reduced to a p-value cut-off of 0.05 without correction for multiple testing. (**B**) AUC values for top-20 TFs identified by both `LLM3D` and MGSI in the yeast metabolic cycle expression clusters. `LLM3D` produces higher AUC values than MGSI, even when the stringency of the latter is reduced to a p-value cut-off 0.05 without correction for multiple testing.

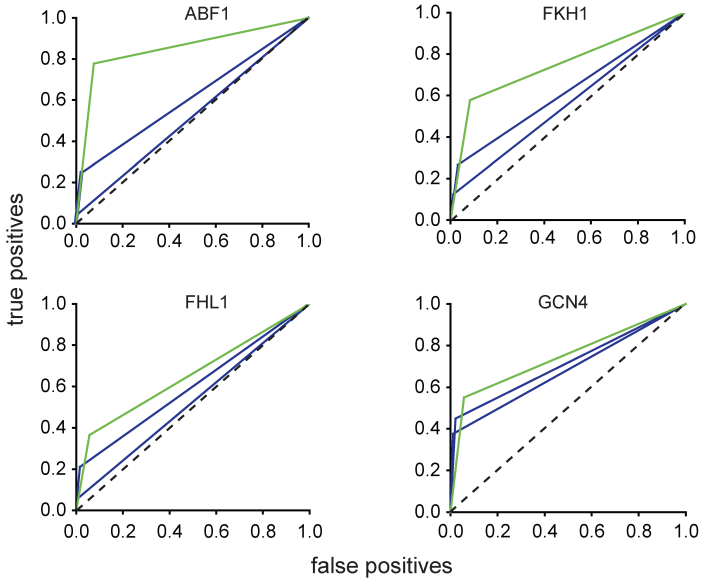

Figure 2.4: Example ROC curves for four TFs showing that `LLM3D` (green curves) indeed produces higher AUC values compared with MGSI (blue curves), indicating that `LLM3D` has better predictive performance.

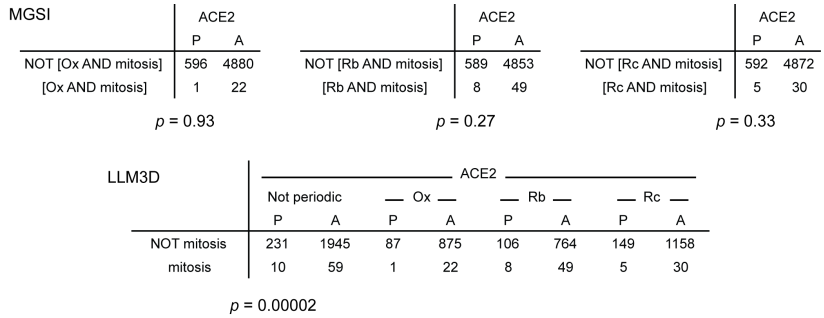

Figure 2.5: Example contingency tables showing that LLM3D detects a significant interaction of ACE2 binding sites with mitosis GO genes in yeast metabolic cycle expression clusters (Ox, Rb and Rc), whereas MGSI does not.

significant TF-target gene associations. On average, LLM3D achieved about a 3-fold increase in the number of true positive TF-target gene predictions compared with MGSI at an equivalent false discovery rate of 10% (Figure 2.3 A). LLM3D consistently performed better, even when the stringency of MGSI was reduced to a p-value cut-off of 0.05 without correction for multiple testing. Importantly, this increase in the number of true positive predictions was not simply the result of an overall increase in TF-target gene predictions, because also the average AUC values are consistently higher (0.74 for LLM3D compared to only 0.59 ($q < 0.1$) or 0.66 ($p < 0.05$) for MGSI; Figure 2.3 B). Example ROC curves for four TFs are plotted in Figure 2.4, and an overview of all 20 TFs tested is provided in Table 2.1. The results reported in Figure 2.3 and in Table 2.1 were obtained using MRM as the repository of true interactions. Similar results were obtained using YEASTRACT (data not shown). In conclusion, LLM3D achieves a significant gain in statistical power and improved predictive performance compared with existing methods with respect to the prediction of TF-target gene interactions.

LLM3D identifies functional transcriptional regulators of the human cell cycle

To test the performance of LLM3D in mammalian systems, we used LLM3D to predict functional TFBSs from temporal gene expression data obtained in synchronized human HeLa cells (Whitfield *et al.* [115]). In this study, more than 500 cell cycle regulated genes were identified, which were classified into the five different cell cycle clusters, i.e., G1/S, S, G2, G2/M and M/G1, comprising 88, 115, 137, 101 and 88 unique genes, respectively. In these five clusters LLM3D identified 63 significantly enriched TFBSs, whereas only 8 TFBSs were detected with MGSI. We used the LLM3D residuals to calculate cluster-specific enrichment of the 36 top-ranking TFBSs in the five top-ranking associated GO classes: DNA metabolic process (GO:0006259), DNA replication (GO:0006260), cell-cycle (GO:0007049), mitosis (GO:0007067), and cell division (GO:0051301)

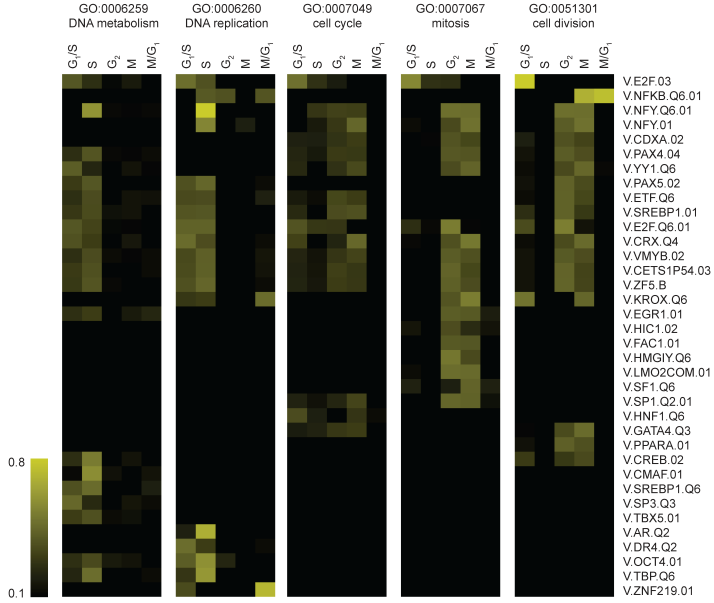

Figure 2.6: Prediction of human cell cycle regulators by LLM3D. Heat maps showing the overrepresentation of 36 TFBSs in each of the 5 expression clusters (G1/S, S, G2, G2/M and M/G1) and in association with the 5 top-ranking cell cycle-related GO classes ('DNA metabolism', 'DNA replication', 'cell cycle', 'mitosis' and 'cell division'). Relative enrichments are indicated as normalized LLM3D residual values. LLM3D predicts many known cell cycle regulators (E2F, SP1, YY1, CREB, NF-Y and EGR1/KROX24), but also many new ones.

(Figure 2.6). As expected, significant TFBS associations with DNA metabolic process and DNA replication genes are primarily detected in the G1/S and S clusters, whereas TFBS associations with mitosis and cell division genes primarily occur in the G2 and G2/M clusters. TFBS associations with cell cycle genes are enriched in all four clusters, indicating that this GO class may provide a more general description of cell cycle genes. Little TFBS enrichment was observed in the M/G1 cluster, suggesting that this cluster may be biologically less informative compared with the other clusters. Importantly, LLM3D identified many previously predicted cell cycle regulators, such as E2F, SP1, YY1, CREB, NF-Y (Elkon *et al.* [28]) and EGR1/KROX24 (Min *et al.* [71]). Due to the lack of genome-wide promoter binding data for most of these TFs, we are unable to formally test LLM3D performance using ROC analysis. Nevertheless, our findings clearly demonstrate that LLM3D is able to predict known regulators of the human cell cycle, and thus validates LLM3D as a method to study mammalian gene regulatory interactions.

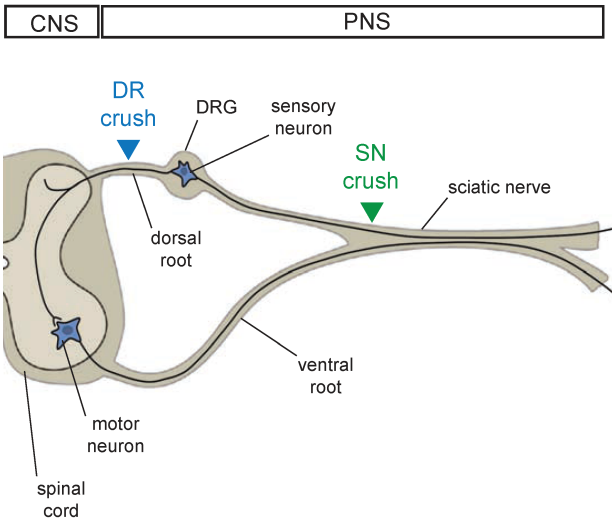

Figure 2.7: Schematic representation of the sensory-motor neuron circuitry and the location of the DRG. A dorsal root crush injures the central projections of DRG sensory neurons, whereas a peripheral nerve crush injures the peripheral projections of the same neurons.

LLM3D identifies novel transcriptional regulators of neuronal injury-induced genes

We next used LLM3D to predict novel transcriptional regulatory interactions underlying neuronal regeneration. We first generated genome-wide expression profiles of dorsal root ganglion (DRG) neurons following nerve damage (Figure 2.7). DRG neurons extend one peripheral axon into the spinal nerve, which regenerates spontaneously after damage, and one central axon into the dorsal root, which has little regenerative capacity. For this study, two groups of animals were subjected either to sciatic nerve (SN) or dorsal root (DR) crush, and at 12, 24, 72 hours and 7 days after the crush, lumbar DRGs L4, L5 and L6 were dissected and total RNA was extracted. Samples were then processed and hybridized to Agilent 44K rat whole-genome arrays. Bayesian Analysis of Time-Series (BATS) (Angelini *et al.* [6]) was used to identify 2,845 genes that are significantly regulated after SN crush and 2,775 genes that are significantly regulated after DR crush relative to control. Out of the 4,735 regulated genes in total, only 885 genes were regulated in both crush paradigms and 3,850 were regulated in a paradigm-specific manner (Figure 2.8), which confirms the notion that SN crush and DR crush elicit very early and distinct gene expression responses in DRG neurons (Stam *et al.* [99]). In line with previous gene expression studies (Costigan *et al.* [22], Schmitt *et al.* [88], Stam *et al.* [99], Szpara *et al.* [103]), we find a strong up regulation of regeneration-associated genes such as Atf3, Pap, Vip, Npy, Gal, Tgm1, Csrp3 and Ankrd1, Gadd45a and Vgf (Figure 2.9).

We separated all 4,735 regulated genes into two distinct expression clusters; one cluster

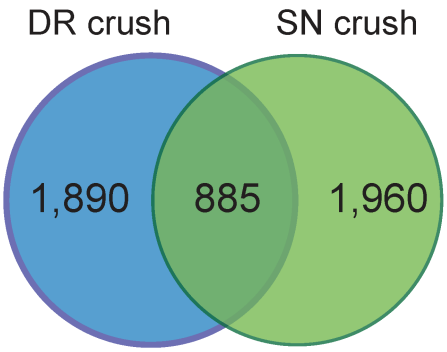

Figure 2.8: Venn diagram showing the number of significantly regulated genes identified in each crush paradigm. The relatively small overlap indicates that SN and DR crush elicit distinct gene responses in DRG neurons.

of genes that are either up-regulated after SN crush or down-regulated after DR crush (named SN>DR) and thus constitutes putative regeneration-promoting genes, and one cluster of genes that are either up-regulated after DR crush or down-regulated after SN crush (named DR>SN) and thus constitutes putative regeneration-inhibiting genes (Figure 2.10**A**). We next used either MGSI or LLM3D to predict transcriptional regulatory interactions underlying gene expression within each of these clusters. Predicted rat TRANSFAC binding sites in the DNA regulatory sequences of all genes present on the array, and informative GO biological process terms were used as input for our analysis (see Section 2.5 for details). Each TFBS-GO pair was then tested for association with the SN>DR and DR>SN gene expression clusters. After correction for multiple testing, MGSI identified only 37 TFBSs and 66 TFBS-GO pairs compared to 157 TFBSs and 1,464 TFBS-GO pairs identified by LLM3D. Because it is unlikely that all these significant TFBS-GO pairs predict biologically equally relevant transcription regulatory interactions, we next used a relative enrichment scoring method to select the 50 TFBSs with the highest expression cluster-specific regulatory potential. This method is based on ranking all predicted TFBSs according to the observed variance in enrichment per TFBS-GO association in each of the two expression clusters, and assumes that more variance is indicative of more expression cluster-specific TFBS-GO associations. Importantly, the 50 TFBSs that were selected include 23 TFBSs that were also detected using MGSI (Figure 2.10**B**). This indicates that LLM3D indeed provides an improvement over MGSI, and extends the number of sites that can be reliably detected.

To further select biologically relevant TFBSs, we also performed LLM3D analysis on the SN>DR and DR>SN expression clusters using only human/mouse/rat (HMR)-conserved TFBSs. Interestingly, one of the most significant TFBS association that was thus identified indicated an overrepresentation of peroxisome proliferator activated receptor γ (PPARγ) binding sites in `neuron differentiation` GO genes in the DR>SN expression cluster (Figure 2.10**C**). In the non-conserved (rat only) binding site analysis, a similar but weaker interaction between PPARα binding sites and `neuron differentiation` genes was

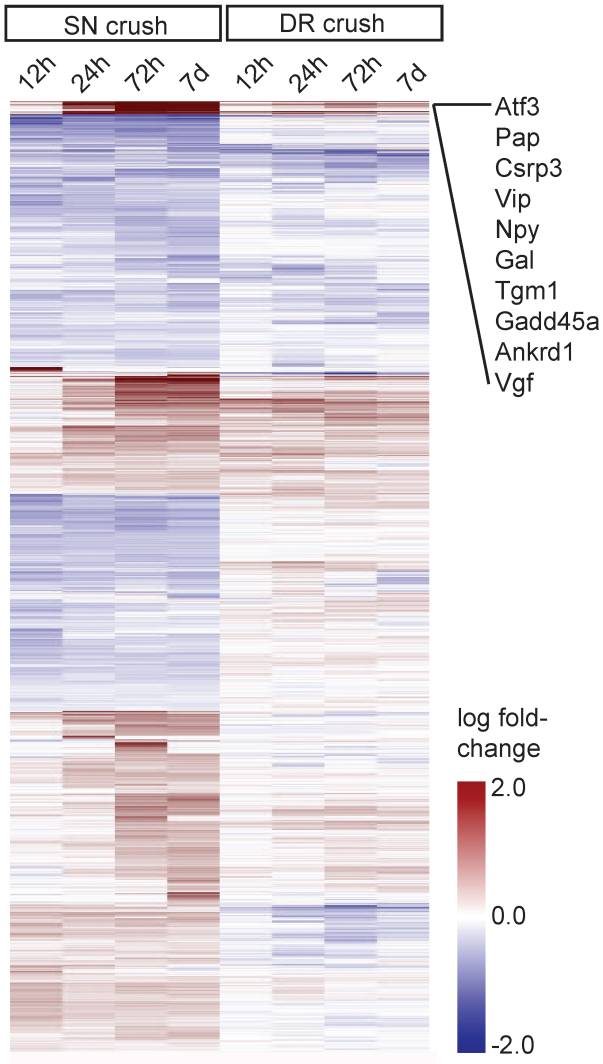

Figure 2.9: Heatmap showing expression profiles of all significantly regulated genes after SN crush.

detected (Figure 2.10**B**). In fact, all PPAR subtypes are reported to recognize the same peroxisome proliferator response element (PPRE: AGGACA-N-AGGACA) (Kliewer *et al.* [55]). We conclude that PPAR binding sites in TRANSFAC differ from each other in informational content, and that focusing on conserved binding sites may help to reduce background for poorly defined binding sites such as PPARγ.

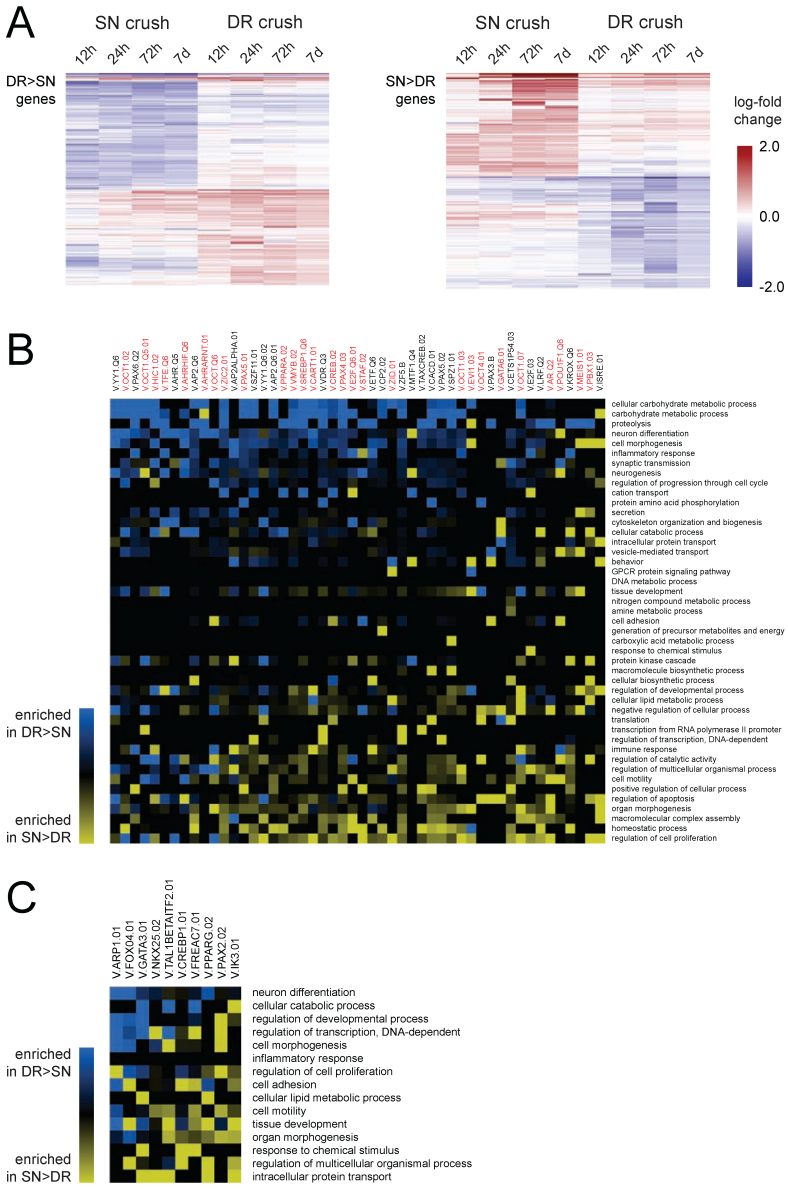

Figure 2.10: Computational prediction of TFBS-GO associations in regeneration-associated genes. (**A**) All genes that are significantly regulated after SN or DR injury were subdivided into two clusters: DR>SN and SN>DR (see text for details). (**B**) Heat map showing the top-50 TFBS-GO associations detected in DR>SN genes (blue) and in SN>DR genes (yellow). TRANSFAC binding sites are on the horizontal axis (binding sites indicated in red were only detected with `LLM3D`, binding sites indicated in black were also identified with MGSI); GO terms are on the vertical axis.(**C**) Heat map as in (**B**), showing the top-10 TFBS with their top-15 GO associations as detected by `LLM3D` using the HMR-conserved TFBSs selection.

Binding	True positives			Area Under Curve		
Site	MGSI.q	MGSI.p	LLM3D	MGSI.q	MGSI.p	LLM3D
ABF1	22	4	70	0.61	0.52	0.85
CBF1	66	34	95	0.76	0.63	0.87
FHL1	20	5	35	0.6	0.53	0.65
FKH1	17	8	37	0.62	0.56	0.75
FKH2	29	13	47	0.66	0.57	0.75
GCN4	49	41	60	0.71	0.68	0.75
HAP1	18	8	32	0.6	0.55	0.67
HSF1	16	15	24	0.69	0.68	0.79
MBP1	50	41	65	0.75	0.71	0.81
MCM1	25	10	35	0.76	0.61	0.86
MSN2	31	26	39	0.68	0.65	0.71
RAP1	16	4	34	0.6	0.53	0.71
REB1	40	0	46	0.65	0.5	0.67
RPN4	22	11	28	0.78	0.64	0.85
SKN7	22	3	27	0.58	0.51	0.59
STE12	19	12	29	0.57	0.55	0.6
SWI4	32	23	43	0.66	0.62	0.71
SWI6	32	28	47	0.64	0.62	0.69
UME6	24	16	56	0.65	0.61	0.86
YAP7	23	15	31	0.63	0.59	0.68

Table 2.1: Comparison of MGSI and LLM3D predictive performance.

2.3 Experimental validation of LLM3D predictions

Because PPAR binding sites were consistently detected in neuron differentiation GO genes using both the non-conserved (rat only) and HMR-conserved prediction methods in LLM3D, we decided to experimentally test whether PPAR TFs are involved in regenerative neurite outgrowth. Most predicted PPAR target genes are expressed in neurons (see Appendix A), which suggests that PPAR TFs may enhance regeneration by regulating the expression of neuron intrinsic regeneration-associated genes. To establish which PPAR subtype might be involved, we tested the effects of different PPAR agonists and antagonists on neurite outgrowth from DRG-like F11 cells and from primary adult DRG neurons. Stimulation of PPARγ, but not PPARα, stimulated neurite outgrowth from primary DRG neurons and from F11 cells, whereas blocking PPARγ, but not PPARα, inhibited neurite outgrowth in both cell types (Figure 2.11(a-d)). These findings show that activation of PPARγ in primary adult DRG neurons, which are the closest to the DRG regeneration paradigm on which our TFBS predictions are based, stimulates neurite outgrowth. Primary DRG cultures are however mixed neuron/glia cultures, and the effects of PPARγ activation or inhibition on DRG neuron outgrowth might be indirectly mediated by glial cells. F11 cell cultures on the other hand are purely neuronal, and the fact that we could replicate our results in F11 cells indicates that PPARγ is a neuron-intrinsic stimulator of neurite outgrowth.

To test whether PPARγ binds directly to the promoters of predicted target genes, we next performed quantitative chromatin immunoprecipitation (ChIP). F11 cells were stimulated with the PPARγ agonist ciglitazone or with DMSO (control) and chromatin complexes were cross-linked after 24 hours and subjected to ChIP using an antibody specific for PPARγ. Immunoprecipitated DNA was then analyzed using quantitative PCR. PCR primers were designed to recognize 100 base pair promoter regions containing the predicted PPRE sites for 9 randomly chosen predicted target genes. As negative controls we used primers recognizing promoter regions of Icer and JunD that lack PPREs. For 8 promoter regions tested, we found a specific interaction with PPARγ, which in most cases was further induced by ciglitazone (Figure 2.11(e))). These findings indicate that LLM3D predicts within a given functional context (i.e., neuron differentiation) PPARγ target gene interactions with an accuracy of more than 80%.

We finally measured the effect of ciglitazone on the expression of the six predicted target genes that show the highest PPARγ-binding. Quantitative PCR measurements indicate that activation of PPARγ with ciglitazone significantly reduces the expression of four of these genes (Figure 2.11(f)), which demonstrates that PPARγ acts as a ligand-dependent repressor of gene expression. Importantly, PPARα agonist Wy-14643 did not affect gene expression levels, nor did any of the pharmacons affect the expression levels of PPARα or PPARγ (Figure 2.11(f)).

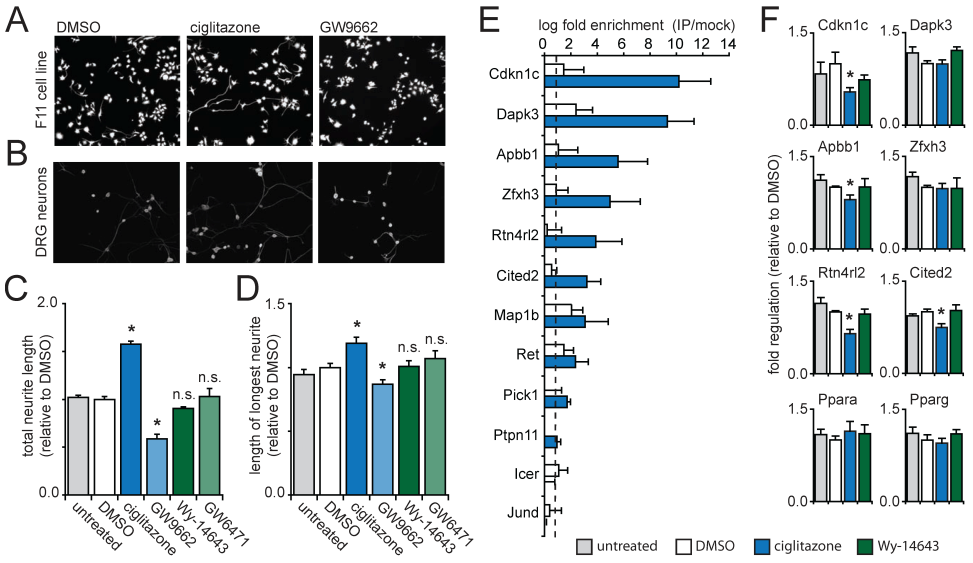

Figure 2.11: Experimental validation of PPARγ binding sites in regeneration-associated genes. (**a**) F11 cells treated with PPARγ agonist ciglitazone show increased neurite outgrowth, whereas cells treated with PPARγ antagonists GW9662 show decreased neurite outgrowth. (**b**) Similar results were obtained for cultured primary adult DRG neurons. (**c**) Quantification of the effects of ciglitazone and GW9662 on F11 cell outgrowth. (**d**) Quantification of the effects of ciglitazone and GW9662 on primary DRG neuron outgrowth. Note that PPARα agonist Wy-14643 and antagonist GW6471 do not affect neurite outgrowth. (**e**) PPARγ binds to the promoters of predicted `neuron differentiation` target genes. Anti-PPARγ immunoprecipitated chromatin from F11 cells treated with DMSO (negative control; white bars) or PPARγ agonist ciglitazone (blue bars) was quantified by PCR using site-specific primers. `Icer` and `JunD` were included as negative control genes. All predicted target genes tested, except `Pick1` and `Ptpn11`, show PPARγ binding above background (dashed line), and for most genes this binding was strongly enhanced by ciglitazone. (**f**) Four out of six PPARγ-binding genes show a significant reduction in expression after ciglitazone treatment (blue bars) compared with DMSO treatment (white bars). PPARα agonist Wy-14643 did not affect gene expression levels (green bars), nor did ciglitazone or Wy-14643 affect the expression levels of PPARγ or PPARα. Bars represent means $\pm$ SD; $* \; p < 0.01$.

2.4 Discussion

Reverse engineering transcriptional regulatory networks from experimental data presents great challenges, particularly in higher organisms. As more genome-wide gene expression and functional data sets become available, there is a growing need for computational methods to analyze these data and accurately infer regulatory relationships from them. Of particular interest are those methods that automatically generate experimentally testable hypotheses regarding the direct regulation of genes by DNA-binding TFs. Combining heterogeneous sources of information, including genome-wide gene expression, DNA sequence and functional annotation, may prove to be essential to accurately predict true regulatory relationships. Here, we present a new method that offers a significant improvement over currently used enrichment based methods, and show that this method can be applied to predict novel, condition specific sets of transcriptional targets in the context of the complexity of the mammalian genome.

The main problem associated with existing methods is that they do not model the joint dependence between gene expression, TFBS presence and gene function. SEA-based methods for instance produce lists in which enriched TFBS and GO terms occur separately. From such lists it is unclear how GO terms and TFBS are jointly related to the gene sets of interest, and thus it is not possible to directly use SEA results to predict functionally homogenous sets of TF target genes. MGSI-based methods on the other hand try to circumvent this problem by using pre-defined GO expression gene sets, and subsequently test these sets for enrichment of TFBSs. Although it makes sense to search for TFBS enrichment in functionally homogeneous sets of co-expressed genes, there are important conceptual problems with this approach that compromise the analysis and adversely affect the power to detect biologically meaningful associations. For instance, MGSI does not really consider gene expression, TFBS presence and GO annotation jointly, but rather collapses gene expression and GO annotation into a single combined variable before computational analysis. Thus, important information about the joint dependence of all three variables is lost. Moreover, by analyzing multiple disjoint gene expression clusters, MGSI aggravates the multiple-testing problem because separate tests are performed for each cluster. `LLM3D` efficiently deals with both problems; it allows modeling of the joint distribution between all variables and reduces the number of tests to be performed.

We validated `LLM3D` performance in two ways. Firstly, we used published yeast gene expression and transcription factor binding data to show that `LLM3D` can indeed detect experimentally validated TFBSs that remain undetected using MGSI. Moreover, analysis of true positive versus false positive rates for multiple TFBSs shows that `LLM3D` predicts TF target genes with higher accuracy. Secondly, we show that `LLM3D` detects both known and novel TFBSs in published human cell cycle gene expression data. Thus, `LLM3D` provides a significant computational improvement for the detection of functional gene regulatory interactions, both in yeast and in mammals.

We next used `LLM3D` to identify novel gene regulatory interactions underlying neuronal regeneration. As expected, `LLM3D` predicted many more significantly enriched TFBSs in regeneration-associated gene sets than MGSI. To be able to focus on the most relevant TFBSs we implemented several selection options. To reduce extensive overlap in GO-defined gene

sets due to the hierarchical structure of the GO tree, we first pre-selected informative GO terms for input into LLM3D. Next, we considered only those LLM3D-predicted TFBSs that showed the highest gene set-specific enrichment. About half of the 50 TFBSs that were thus obtained remained undetected using MGSI. Finally, we also implemented an option in LLM3D to limit the search to HMR-conserved TFBSs, which may further improve detection of biologically relevant TFBS-GO associations.

One of the most significant TFBS-GO interactions that was detected with LLM3D, but not with MGSI, predicts a role for PPAR transcription factors in the regulation of genes that are involved in neuron differentiation. This finding raised our interest because activation of PPARγ in spinal cord injury models has beneficial effects on the functional outcome (McTigue *et al.* [69]; Park *et al.* [76]), but it is not clear whether these effects are directly on the damaged neurons, or whether PPARγ reduces the secondary inflammatory response (McTigue [70]). Our findings show that PPARγ, but not PPARα, stimulates neurite outgrowth of DRG neurons. Moreover, this effect of PPARγ is neuron-intrinsic since we also observe it in DRG-like F11 cells, which in the presence of Forskolin acquire a neuronal phenotype. Activated PPARγ binds to promoters of predicted target genes and reduces their expression. Importantly, several predicted PPARγ target genes are known inhibitors of neurite outgrowth (e.g. Rtn4rl2, Slit1, Hes5; see Appendix A), which suggests that PPARγ promotes neurite outgrowth by repressing growth-inhibitory genes. At this moment we can only speculate about the relevance of these findings for neuronal regeneration *in vivo*. The primary ligands of PPARγ are polyunsaturated fatty acids (Hihi *et al.* [47]). Following nerve crush and degeneration of the myelin sheath, free myelin lipids are taken up by macrophages and released again as fatty acids to be incorporated into the newly forming myelin sheath (Goodrum *et al.* [39]). We propose that injured axons might benefit from fatty acid production in the damaged nerve, and that the neuron-intrinsic lipid sensing properties of PPARγ may play an important role in conveying injury signals from the crush site to the nucleus. This hypothesis is supported by several reports showing beneficial effects of fatty acids on neurite outgrowth *in vitro* (Liu *et al.* [63]; Robson *et al.* [84]) and on neuronal regeneration *in vivo* (McTigue *et al.* [69]; Park *et al.* [76]), and the induction of fatty acid-binding proteins in regenerating axons (De Leon *et al.* [61]).

One of the challenges left unaddressed in the current implementation of our method is that transcriptional regulation in higher organisms is believed to be highly combinatorial, and that the spatio-temporal expression of genes is influenced by multiple regulatory TFs that form complexes at multiple TFBSs. Although some basic models for the cooperative effect of multiple TFs on the expression of target genes have been suggested [5, 59, 93, 112], in general the *cis*-regulatory grammar underlying gene regulation is still poorly understood. Moreover, combinatorial models of gene regulation are difficult to validate and the effect of different TFs on target genes is therefore most often studied independently. As soon as reliable and genome-wide descriptions of *cis*-regulatory modules become available we will adapt LLM3D to allow modeling of *cis*-regulatory modules in addition to individual TFBSs.

In conclusion, LLM3D provides an important improvement over existing computational methods in identifying functional TFBSs from gene expression data. Its unique property of testing the joint association between multiple features (e.g., gene expression, gene function and TFBS occurrence) based on one table allows further generalization to tables with more

dimensions including additional relevant gene attributes. The implementation of such multi-dimensional computational methods will be of critical importance in order to extract biologically meaningful information from the increasing number, size and diversity of data sets generated by biologists.

2.5 Methods

LLM3D—description

For each TFBS-GO pair of interest, LLM3D cross-classifies all genes according to GO annotation, TFBS presence, and gene expression to obtain a three dimensional contingency table. The rows of this table, indexed by i, correspond to the GO variable which has two categories. A gene is classified in the first category ($i = 1$) if it is *not* annotated to the GO term under consideration, and in the second ($i = 2$) if it *is*. The columns of the table, indexed by j, correspond to the TFBS variable which also has two categories. A gene is classified in the first category ($j = 1$) if the gene has *no* TFBS for the TF under consideration, and in the second ($j = 2$) if it *does*. The third dimension of the table, the layers indexed by k, correspond to the observed gene expression. This variable has K categories, as many as the number of gene expression clusters. A gene is categorized in the k-th layer if it belongs to the k-th cluster, $k = 1, \ldots, K$. Note that since LLM3D is designed to do a genome-wide analysis, the genes in the cluster that corresponds to $k = 1$ are supposed to represent a "background" or "reference" set of genes that are not regulated/expressed in the gene expression experiment under consideration. The main statistical analysis of LLM3D consists of finding a good model that describes the observed counts in the three dimensional table.

To define the class of models that LLM3D considers, we introduce the following notation. Let n_{ijk} denote the observed number of genes in row i, column j and layer k. Furthermore, let the summation over an index be denoted by a dot in the subscript at the position of that index. Then the marginal totals of the table are given by

$$n_{ij.} = \sum_{k=1}^{K} n_{ijk}, \qquad n_{i.k} = \sum_{j=1}^{2} n_{ijk}, \qquad n_{.jk} = \sum_{i=1}^{2} n_{ijk},$$

$$n_{i..} = \sum_{j=1}^{2}\sum_{k=1}^{K} n_{ijk} = \sum_{j=1}^{2} n_{ij.} = \sum_{k=1}^{K} n_{i.k},$$

$$n_{.j.} = \sum_{i=1}^{2}\sum_{k=1}^{K} n_{ijk}, \qquad n_{..k} = \sum_{i=1}^{2}\sum_{j=1}^{2} n_{ijk},$$

and finally, the grand total is given by

$$n_{...} = \sum_{i=1}^{2}\sum_{j=1}^{2}\sum_{k=1}^{K} n_{ijk}.$$

It is assumed that the observed counts n_{ijk} are realizations of random variables N_{ijk}. Let π_{ijk} denote the probability that a gene falls in row i, column j and layer k, and let $m_{ijk} = \mathbb{E}(N_{ijk})$ be the expected number of genes in row i, column j and layer k. The $2 \times 2 \times K$-vector of all cell counts $N_{\text{counts}} = (N_{111}, \ldots, N_{22K})$ is assumed to have a multinomial distribution with parameters $(n_{...}, \pi_{111}, \ldots, \pi_{22K})$, where the sum of all π_{ijk}s is equal to 1. Thus, if we

observed a sample of $n_{...} = N$ genes, the contingency table classifies the N genes of the sample into $2 \times 2 \times K$ sub-populations, and the sizes of these subpopulations are determined by this multinomial distribution. The expected number of genes in sub-population (i, j, k) is $m_{ijk} = N\pi_{ijk}$. Now whether or not a gene is classified in sub-population (i, j, k) depends on the dependence structure of the factors that define the rows, columns and layers, that is, on the (in)dependence structure of the three variables GO annotation, TFBS presence, and gene expression. The different types of (in)dependence structure can be expressed in terms of different additional restrictions on the π_{ijk}, and each type of (in)dependence structure thus is described by a different model. To find the most suited of these models, `LLM3D` first tests whether the model for complete independence between the factors holds. If this test shows strong evidence against independence of the factors, `LLM3D` then selects a good alternative model out of all possible dependence models.

The model that assumes that GO annotation, TFBS presence, and gene expression are mutually independent is referred to as $M^{(0)}$, and the corresponding additional restrictions on the probabilities π_{ijk} for this model are

$$M^{(0)}: \ \pi_{ijk} = \pi_{i..}\pi_{.j.}\pi_{..k}, \quad i = 1,2 \quad j = 1,2 \quad k = 1,\ldots,K. \tag{2.1}$$

We note that this can also be written as

$$M^{(0)}: \ N\pi_{ijk} = N\pi_{i..}\pi_{.j.}\pi_{..k}, \quad i = 1,2 \quad j = 1,2 \quad k = 1,\ldots,K.$$

or equivalently, taking logarithms on both sides,

$$M^{(0)}: \log(m_{ijk}) = \eta + \alpha_i + \beta_j + \gamma_k, \quad i = 1,2 \quad j = 1,2 \quad k = 1,\ldots,K,$$

hence the name *loglinear model*. To test the null hypothesis of independence—or equivalently the null hypothesis that $M^{(0)}$ holds—against the alternative of dependence a likelihood ratio test is performed in which the likelihood under the model $M^{(0)}$ is compared to a model with no additional restrictions on the parameters π_{ijk}. The latter model is called the saturated model and we denote it by $M^{(S)}$. It has a parameter for every cell in the table, hence its name. Because the probabilities π_{ijk} add up to one, the model's number of free parameters is equal to $4K - 1$, the number of cells in the table-1, whereas the model $M^{(0)}$ has, due to the additional restrictions in (2.1), only $K + 1$ free parameters. In the test statistic the unknown values of the π_{ijk} are replaced by their maximum likelihood estimates. Under model $M^{(0)}$ the maximum likelihood estimate $\hat{\pi}_{ijk}^{M^{(0)}}$ of π_{ijk} is given by

$$\begin{aligned} \hat{\pi}_{ijk}^{M^{(0)}} &= \hat{\pi}_{i..}\hat{\pi}_{.j.}\hat{\pi}_{..k} \\ &= (n_{i..}/N)(n_{.j.}/N)(n_{..k}/N), \end{aligned}$$

and under $M^{(S)}$ the maximum likelihood estimate is $\hat{\pi}_{ijk}^{M^{(S)}}$ and satisfies

$$\hat{\pi}_{ijk}^{M^{(S)}} = \hat{\pi}_{ijk} = n_{ijk}/N.$$

Under the null hypothesis that the independence model $M^{(0)}$ holds, and when the number of counts is large, the likelihood ratio test statistic is approximately distributed as a chi-square distribution with number of degrees of freedom equal to the difference of the numbers of estimated parameters between the two models $M^{(S)}$ and $M^{(0)}$, namely $4K - 1 - (K + 1) = 3K - 2$. The observed value of the likelihood ratio test statistic for the contingency table is

$$G^2_{M^{(0)}} = 2 \sum_{i=1}^{2} \sum_{j=1}^{2} \sum_{k=1}^{K} n_{ijk} \log \left(\frac{n_{ijk}}{\hat{m}^{M^{(0)}}_{ijk}} \right), \tag{2.2}$$

where $\hat{m}^{M^{(0)}}_{ijk}$ is the estimated expected number of genes in row i, column j and layer k under $M^{(0)}$ given by

$$\begin{aligned} \hat{m}^{M^{(0)}}_{ijk} &= N \hat{\pi}^{M^{(0)}}_{ijk} \\ &= n_{i..} n_{.j.} n_{..k} / (N^2). \end{aligned}$$

Hence, the p-value $p^{M^{(0)}}$ for the likelihood ratio test is obtained by

$$p^{M^{(0)}} = \mathbb{P}(\chi^2_{3K-2} \geq G^2_{M^{(0)}}), \tag{2.3}$$

where χ^2_{3K-2} is a chi-square distributed random variable with $3K - 2$ degrees of freedom. `LLM3D` rejects the null hypothesis of mutual independence between rows, columns and layers of the contingency table at significance level α if

$$p^{M^{(0)}} < \alpha. \tag{2.4}$$

Next, if the null hypothesis of independence is rejected, `LLM3D` selects a good alternative model out of the possible *dependence* models. For a two-dimensional contingency table there are only two models of interest: one in which the rows and columns are independent and one in which they are not. For a three-dimensional contingency table besides the independence model there are seven natural dependence models to consider. These models differ in the restrictions they put on the probabilities π_{ijk}, or, equivalently, in the number of free parameters used to model the expected counts in the cells of the table, and in the dependence relationships they imply between the rows, columns and layers of the table. These seven different models are listed in Table 2.2. The model $M^{(0)}$, which assumes mutual independence between the rows, columns and layers of the table already has been discussed above. Models $M^{(1)}$, $M^{(2)}$, $M^{(3)}$ are models in which one factor is independent of the other two. For instance, model $M^{(3)}$ assumes that the factor gene expression is independent of GO en TFBS, but that GO en TFBS may be dependent. Models $M^{(4)}$, $M^{(5)}$ and $M^{(6)}$ are so-called conditional independence models, i.e. models in which one factor is independent of a second factor *given* the third factor. Model $M^{(7)}$ does not have a simple interpretation in terms of independence. In this model there is association between all pairs of variables and the association between any two of the variables is the same at all levels of the third. Its

Model	Description	Restrictions on π_{ijk}
$M^{(0)}$	Mutual independence	$\pi_{ijk} = \pi_{i..}\pi_{.j.}\pi_{..k}$
$M^{(1)}$	GO $\perp$ GE,BS	$\pi_{ijk} = \pi_{i..}\pi_{.jk}$
$M^{(2)}$	BS $\perp$ GE,GO	$\pi_{ijk} = \pi_{.j.}\pi_{i.k}$
$M^{(3)}$	GE $\perp$ GO,BS	$\pi_{ijk} = \pi_{..k}\pi_{ij.}$
$M^{(4)}$	BS $\perp$ GO|GE	$\pi_{ijk} = \pi_{i.k}\pi_{.jk}/\pi_{..k}$
$M^{(5)}$	GE $\perp$ GO|BS	$\pi_{ijk} = \pi_{ij.}\pi_{.jk}/\pi_{.j.}$
$M^{(6)}$	GE $\perp$ BS|GO	$\pi_{ijk} = \pi_{ij.}\pi_{i.k}/\pi_{i..}$
$M^{(7)}$	uniform association	see text
$M^{(S)}$	saturated model	no restrictions on π_{ijk}

Table 2.2: Overview of models fitted by `LLM3D`.

restrictions on the π_{ijk} are in the form of the equality of the following odds-ratios

$$M^{(7)} : \frac{\pi_{111}\pi_{2j1}}{\pi_{211}\pi_{1j1}} = \frac{\pi_{11k}\pi_{2jk}}{\pi_{21k}\pi_{1jk}},$$

for $j = 1,2$ and $k = 1,\ldots,K$. See [19] for more details. The final model is the earlier introduced saturated model $M^{(S)}$, which has no restrictions on the probabilities π_{ijk}.

For each model $M \in \mathscr{M} = \{M^{(1)},\ldots,M^{(S)}\}$, estimates $\hat{\pi}^M_{ijk}$ of π_{ijk} can be computed by maximum likelihood under the assumption that the model M holds, and $\hat{m}^M_{ijk} = n_{...}\hat{\pi}^M_{ijk}$ is the estimated expected number of genes in row i, column j and layer k under model M. For each model $M \in \mathscr{M}$, the value of the corresponding likelihood ratio test statistic then can be computed analogously to (2.2):

$$G^2_M = 2\sum_{i=1}^{2}\sum_{j=1}^{2}\sum_{k=1}^{K} n_{ijk}\log\left(\frac{n_{ijk}}{\hat{m}^M_{ijk}}\right).$$

Under model M this G^2_M statistic is again approximately chi-square distributed when the number of observations is large, but with a different number of degrees of freedom, namely ($4K-1$−the number of estimated parameters). To select a good dependence model for the table, we could in principle compare the models $M \in \mathscr{M}$ using G^2_M, since a large value of G^2_M would indicate lack of fit of model M. However, the models in $\mathscr{M}$ differ in the number of free parameters used to explain the observed counts and an increase in the number of parameters will result in a decrease in G^2_M. As a result, G^2_M itself is not appropriate for model selection. Instead, we will use Akaike's information criterion (AIC) [3] for model selection, which can be calculated directly from the G^2_M statistic. Finding the best model M^{AIC} in $\mathscr{M}$ based on the AIC criterion amounts to the following minimization (see [19, (Christensen

1997)])

$$\begin{aligned} M^{AIC} &= \arg\min_{M\in\mathscr{M}} A_M = \arg\min_{M\in\mathscr{M}}\{G^2_M - (r_{M^{(S)}} - 2r_M)\} \\ &= \arg\min_{M\in\mathscr{M}} A_M - r_{M^{(S)}} = \arg\min_{M\in\mathscr{M}}\{G^2_M - 2(r_{M^{(S)}} - r_M)\} \\ &= \arg\min_{M\in\mathscr{M}}\{G^2_M - 2df_M\}, \end{aligned} \tag{2.5}$$

where r_M is the number of degrees of freedom, or number of estimated parameters, of model M, and df_M is the number of degrees of freedom of the test statistic G^2_M.

Example 2.5.1

Let us consider an example using real experimental data. We consider the GO term `cell cycle` and the TRANSFAC TFBS with id `V.E2F.Q6.01`. The row variable `cell cycle` classifies genes according to whether they are known to be involved in the cell cycle ($i = 2$), or not ($i = 1$). The column variable `V.E2F.Q6.01` classifies genes according to whether they have an `E2F` binding site ($j = 2$ indicating presence (P)) or they do not ($j = 1$ indicating absence (A)). The gene expression variable is defined using the clustering of the rat DRG gene expression data. Note that for this example $K = 3$. There is a cluster of "reference" genes ($k = 1$), containing genes not significantly regulated in the gene expression experiment) and two clusters of regulated genes, referred to as DR > SN ($k = 2$) and SN > DR ($k = 3$). The observed table of counts is presented in Table 2.3.

		Gene expression clusters					
		Ref		DR > SN		SN > DR	
		$k = 1$		$k = 2$		$k = 3$	
	E2F TFBS	A	P	A	P	A	P
		$j = 1$	$j = 2$	$j = 1$	$j = 2$	$j = 1$	$j = 2$
GO term	$i = 1$	7100	333	872	48	821	51
`cell cycle`	$i = 2$	321	27	61	8	52	5

Table 2.3: Observed gene counts in the DRG SN/DR experiments.

For this table `LLM3D` fits all models described in Table 2.2 and computes the G^2_M statistic for each of them. Table 2.4 lists the values of the statistics on which the `LLM3D` model selection is based.

In this example, $p^{M^{(0)}} < 1 \times 10^{-5}$, so that the null hypothesis of independence is rejected. Of the dependence models $M^{(7)}$ is selected as M^{AIC}, as it has the lowest value of $G^2_M - df_M$ of

Model	G^2_M	df_M	$G^2_M - 2df_M$
$M^{(0)}$	30.354	7	16.354
$M^{(1)}$	25.559	5	15.559
$M^{(2)}$	16.253	5	6.253
$M^{(3)}$	18.858	6	6.858
$M^{(4)}$	11.459	3	5.459
$M^{(5)}$	14.064	4	6.064
$M^{(6)}$	4.758	4	-3.242
$M^{(7)}$	0.534	2	-3.466
$M^{(S)}$	0	0	0

Table 2.4: Model selection for data in Table 2.3.

the considered models.
***

Finally, for each TFBS-GO pair and corresponding table T, LLM3D records the obtained $p_T^{M^{(0)}}$ and M_T^{AIC}. A Benjamini Hochberg FDR correction [12] for multiple testing is used to convert each $p_T^{M^{(0)}}$ into a multiple testing corrected $q_T^{M^{(0)}}$. The user then may select a subset $\mathscr{T}$ of tables (TFBS-GO pairs) of interest, based on the values $q_T^{M^{(0)}}$ and M_T^{AIC}. For all tables $T \in \mathscr{T}$ and for each gene expression cluster k of interest, the enrichment e_k^T of genes in the cell in row 2, column 2 and layer k is measured as

$$e_k^T = \frac{n_{22k}^T - \hat{m}_{22k}^{T,M^{(0)}}}{\sqrt{\hat{m}_{22k}^{T,M^{(0)}}}}.$$

This enrichment measure is then used to compare the enrichment for the TFBS-GO pairs of interest in different gene expression clusters and to predict context specific functional targets of TFs.

LLM3D—analysis

In the analysis of the yeast data, we selected informative GO terms at level 20. For every TFBS-GO pair, we cross-classified the genes into a three-dimensional table. For every table T we ran the LLM3D analysis, recorded $p_T^{M^{(0)}}$ and M_t^{AIC}, and converted $p_T^{M^{(0)}}$ into $q_T^{M^{(0)}}$ as described above. Then we selected the subset of tables

$$\mathscr{T} = \{T : q_T^{M^{(0)}} < 0.1 \quad \text{AND} \quad M_T^{AIC} \in \{M^{(4)}, M^{(5)}, M^{(6)}, M^{(7)}, M^{(S)}\}\},$$

because the models $M^{(4)}, M^{(5)}, M^{(6)}, M^{(7)}, M^{(S)}$ all imply enrichment of TFBSs in genes that share GO annotation and/or expression cluster membership. For the analysis of yeast data, we demonstrated that the observed enrichment is predictive of functionality of the binding sites, i.e. that the genes in the clusters with positive enrichment are indeed known targets of the corresponding TF. For $T \in \mathscr{T}$, we predicted as targets of the TF corresponding to the TFBS of table T all genes in all cells $(2, 2, k)$, for which $e_k^T > 0$, for $k = 2, 3, 4$.

In the application of LLM3D to human cell cycle and rat DRG gene expression data, we selected a reduced subset

$$\mathscr{T} = \{T : q_T^{M^{(0)}} < 0.1 \quad \text{AND} \quad M_T^{AIC} \in \{M^{(7)}, M^{(S)}\}\},$$

of tables (TFBS-GO pairs) for further analysis, because we believe that the dependence implied by models $M^{(7)}$ and $M^{(S)}$ are most predictive of the *differences* between the clusters.

LLM3D—visualization of results

To visualize the results of an LLM3D analysis, we plot two-dimensional heatmaps in which rows represent GO terms and columns TFBSs. Hence, each little square in the heatmap represents a single TFBS-GO pair and, with it, its corresponding LLM3D table. In the heatmap, we compare the enrichment of TFBSs in two different clusters as follows. For a TFBS-GO pair corresponding to table T and two clusters k_1 and k_2, we compute the enrichment in k_1 *relative to* k_2 using the score

$$s_T = \begin{cases} 0, & \text{if} \quad e_{k_1}^T < 0 \quad \text{and} \quad e_{k_2}^T > 0 \\ 0.5, & \text{if} \quad e_{k_1}^T < 0 \quad \text{and} \quad e_{k_2}^T < 0 \\ e_{k_1}^T / (e_{k_1}^T + e_{k_2}^T), & \text{if} \quad e_{k_1}^T > 0 \quad \text{and} \quad e_{k_2}^T > 0 \\ 1, & \text{if} \quad e_{k_1}^T > 0 \quad \text{and} \quad e_{k_2}^T < 0 \end{cases}.$$

The relative enrichment score s_T computed in this way is a measure ranging from 0 to 1 indicating in which of the two clusters the enrichment of the TFBS is most prominent. If there is no positive enrichment in any of the two clusters, s_T equals 0.5. For squares in the heatmap corresponding to TFBS-GO tables T, for which $T \notin \mathscr{T}$, we also set $s_T = 0.5$. The value 0.5 represents no enrichment and squares with this value receive a neutral color. For $s_T \in [0, 0.5)$ and $s_T \in (0.5, 1]$ we choose two different colors to represent enrichment in the two different clusters. In case not all TFBS-GO pairs can be represented in the heatmap, the most relevant part is selected for visualization as follows. To select which results are of biological interest, all TFBSs and GO terms predicted to be significantly associated with one or more gene expression clusters by LLM3D, are ranked according to the sample variance of their enrichment scores s_T over all associated GO terms and TFBSs, respectively. The highest ranking ones are included in the heatmap.

MGSI

For a given gene expression cluster and GO term, MGSI first generates a new gene set by intersecting the genes in the expression cluster with the set of genes annotated to the GO term. Enrichment of any TFBS in this new set is tested in the classical way using a Fisher's exact test (one-sided) for two-dimensional contingency tables. Because gene cluster-GO intersections are tested for enrichment of many different TFBSs, a Benjamini Hochberg correction is applied to the resulting p-values to correct for multiple testing with the aim of keeping the false discovery rate (FDR) at 10%. In the reanalysis of the yeast metabolic cycle data, we also present results based on using the original p-values without correction for multiple testing. When significant enrichment of TFBSs in a gene expression-GO set is found, the genes in this set with a predicted TFBS are predicted to be targets of the corresponding TF.

Yeast TFBS annotation

Yeast ORF sequences with introns and untranslated regions 1,000 bp immediately upstream of the initial ATG were downloaded from the Saccharomyces Genome Database (SGD) on `http://www.yeastgenome.org`. Log-odds matrices representing PSSM models for binding sites were downloaded from `http://fraenkel.mit.edu/improved_map/` and converted to probability matrices to be used with the `Motifscanner` tool (Aerts *et al.* [1]). `Motifscanner` was used to computationally predict binding sites for all TFs on both DNA strands with the "prior probability" parameter set to 0.15. We generated a 3^{rd} order Markov background model trained on the SGD sequences with the accompanying `CreateBackgroundModel` tool.

Mammalian TFBS annotation

Gene regulatory sequences (5,000 bp upstream to 2,000 bp downstream of the predicted transcription start site) for all human, mouse and rat genes identifiable by Entrez Gene ID were downloaded using the `biomaRt` package under `R`. Potential TFBSs were predicted in silico using all 214 vertebrate non-redundant position weight matrices in the TRANSFAC Professional database (release 11.1) (Matys *et al.* [68]) and the supplied MATCH-tool (Kel *et al.* [54]) with parameters set to minimize false positives. The MATCH output was used to create a binary matrix with rows corresponding to regulatory sequences and columns corresponding to TRANSFAC matrices. In this matrix, Ś1Š represents the presence of at least one predicted TFBS, whereas Ś0Š represents the absence of predicted TFBSs. In addition, all human, mouse and rat genes in `LLM3D` were also annotated with human/mouse/rat (HMR) conserved TFBSs downloaded from `http://genome.ucsc.edu/`. This allows `LLM3D` analysis to be limited to evolutionary conserved binding sites only.

GO pre-selection

Yeast GO annotation data were extracted from the R-package `org.Sc.sgd.db`, which was downloaded from `http://www.bioconductor.org`. GO Biological Process annotations for human, mouse and rat genes were retrieved from `http://www.geneontology.org/`. Informative GO terms were selected as follows. For any GO term i, let GO(i) be the set of genes whose annotation contains term i and let N(i) be the size of that set. We let Child(i) denote the set of children of i in the directed acyclic Gene Ontology graph. Let M(i) be the maximum over N(r), for terms r in Child(i). For any positive number γ and any term i, we now say that i is the *most informative* GO term *at level* γ if N(i) $> \gamma$ and M(i) $< \gamma$. For reanalysis of the yeast metabolic cycle data and the human cell cycle data we considered all most informative GO terms at level 20 in the corresponding gene expression clusters. For the analysis of the neuronal regeneration data we selected most informative GO terms at level 50.

Yeast metabolic cycle data

For reanalysis of the yeast metabolic cycle expression data, we used the original clustering from Tu *et al.* [108]. The MRM refined regulatory map providing true interactions between TFs and target genes based on ChIP-chip data from MacIsaac *et al.* [66] was downloaded from `http://fraenkel.mit.edu/improved_map`. True TF-target gene interactions reported in the YEASTRACT database [104] were downloaded from `http://www.yeastract.com`. For validation of predicted regulatory interactions we used a `RegulationMatrix` containing all documented regulatory interactions in either MRM or YEASTRACT. Yeast GO annotation data were extracted from the R-package `org.Sc.sgd.db`, which was downloaded from `http://www.bioconductor.org`.

Human cell cycle data

`LLM3D` was used to reanalyze the human cell cycle gene expression dataset published by Whitfield *et al.* [115]. The original gene clusters were downloaded from `http://genome-www.stanford.edu/Human-CellCycle/HeLa/`.

Animals and surgical procedures

Adult Wistar rats (220 g; Harlan, The Netherlands) were subjected to either sciatic nerve or dorsal root crush as described previously (Stam *et al.* [99]) in approval with the KNAW animal experimentation committee for animal well fare. L4-6 DRGs were isolated at 12 h, 24 h, 72 h, and 7 days after surgery. Control DRGs were obtained from three uninjured animals.

Microarray hybridization, normalization and analysis

Total RNA was isolated from L4, 5 and 6 DRGs using Trizol (Invitrogen; Carlsbad, CA). RNA pooled from three control animals served as a common reference sample. RNA samples were

amplified, labeled and hybridized to Agilent 44K Rat Whole-Genome expression arrays using standard Agilent protocols (Agilent; Santa Clara, CA). Arrays were scanned using an Agilent scanner and data were read using Agilent Feature Extraction software. Array data were further processed using the R packages Bioconductor (Gentleman *et al.* [36]) and limma (Linear Models for Microarray Data, Ritchie *et al.* [83]) for standard background subtraction and loess normalization. For statistical analyses we used the Bayesian approach for microarray time-course data developed by Angelini *et al.* [6]. This algorithm is implemented in a Matlab executive, termed Bayes Analysis of Time-Series (BATS). Heatmaps and hierarchical clusters were generated using TIGR MeV software (http://www.tm4.org/mev.html).

Expression clusters

The log fold gene expression change (relative to control; averaged over three replicates per time point) in both experiments (SN and DR crush) was calculated for each gene. Expression data from significantly regulated genes following SN and DR crush were analyzed separately using a standard principal component analysis algorithm under R. For each gene, we used the coefficient corresponding to the first principal component to further define two homogeneous gene expression clusters: one containing genes that are either upregulated after DR crush or downregulated after SN crush (DR>SN), and one containing genes that are either upregulated after SN crush or downregulated after DR crush (SN>DR). For a small group of genes that were significantly regulated following both crushes and for which the dominant direction of log fold change (i.e. either up- or downregulation) coincided in both experiments, we compared the average log fold change of expression following SN and DR crush directly for classification into (DR>SN) or (SN>DR).

Cell culture and transfections

F11 cells were maintained as previously described (Stam *et al.* [99]). For pharmacological stimulations F11 cells were plated in 96-well plates. Medium was replaced with DMEM containing 0.5% FCS and the desired concentration of PPAR agonists (ciglitazone for PPARγ and Wy-14643 for PPARα) or antagonists (GW9662 for PPARγ and GW6471 for PPARα; all from Sigma-Aldrich, St. Louis, MO) was added. Cells were fixed two days later and stained with anti-beta-III-tubulin (Sigma-Aldrich). Neurite outgrowth was quantified using a Cellomics KineticScan HCS Reader and the Neuronal Profiling 3.5 Bioapplication (Cellomics Inc., Pittsburgh, PA, USA). Per well 500-1,000 cells were analyzed and neurite total length per cell was calculated. Dissection and dissociation of primary adult DRG neurons was carried out as described13. After 40 hours in culture neurons were fixed and immunostained with anti-beta-III-tubulin. The longest neurite of each of 100-200 neurons was measured using the ImageJ Simple Neurite Tracer plugin.

Chromatin immunoprecipitation (ChIP) and quantitative (RT-)PCR analysis

F11 cells were plated in 15 cm plates, and stimulated with 10 μM forskolin and 10 μM ciglitazone or DMSO as control for 24 hours. Chromatin of F11 cells was then cross-linked with 1% formaldehyde for 10 minutes and subsequently quenched with 125 mM glycine for 5 minutes. Cells were washed with cold PBS, nuclei were extracted with cell lysis buffer (10 mM EDTA, 10 mM HEPES, 0.25% Triton X-100) and lysed with SDS lysis buffer (1% SDS, 10 mM EDTA in 20 mM Tris-HCl). Cross-linked chromatin was sheared with 4 pulses of 15 sec yielding products of 200-1,000 bp in length. Immunoprecipitation was performed with anti-PPARγ (H-100, Santa Cruz Biotechnology) overnight with rotation at 4 žC. Immuno-complexes were then captured with protein A/G beads (Santa Cruz Biotechnology) pre-incubated with sonicated salmon sperm DNA. Complexes were washed and eluted with elution buffer (1% SDS, 100 mM NaHCO3). The eluates were proteinase K treated (215μg/ml) and incubated at 65 žC for overnight. DNA was purified by phenol/chloroform isolation and subsequent ethanol precipitation. Quantitative PCR was performed using site-specific primers in duplicate on a Roche LightCycler with 2x SYBR green ready reaction mix (Applied Biosystems). Normalized enrichment values were calculated by subtracting the Ct value of the IP sample from the Ct value of the mock IP samples, each normalized to the input sample. Promoter regions with $>$1.5 log fold enrichment were considered as true targets. For gene expression level measurements, RNA was isolated from F11 cells using Trizol and reverse-transcribed into cDNA as previously described (Stam *et al.* [99]). Ct values were normalized to Gapdh and Nse as reference genes. Fold changes were calculated relative to DMSO-treated cells. Specificity of all primers was checked by visual inspection of dissociation curves.

Software availability

We developed an `LLM3D` `R`-package which is freely available upon request. A description of the functions in this package can be found in Appendix B.

Acknowledgements

This work received financial support from the Netherlands Organization for Scientific Research (NWO; CLS grant 635.100.008), from the Dutch Ministry of Economic Affairs (SenterNovem grant ISO52022), and from the Center for Medical Systems Biology (CMSB) in the framework of the Netherlands Genomics Initiative (NGI).

THREE

GEMULA

Regression models, in which predictor variables represent TFBSs, can be used to identify associations between *cis*-regulatory DNA motifs occurring in gene promoter sequences and observed variation in gene expression. The TFBS motifs can be represented in several ways that give rise to different predictors. In this chapter we compare regression based approaches for modeling of gene expression data that use different types of regression models and different ways of representing TFBSs. We show that linear models can be used to model synergistic interactions between predictors. We propose `GEMULA`, a strategy based on linear models that is fast, considers a wide range of biologically plausible models and selects parsimonious and interpretable models from experimental data. On yeast gene expression data, we show that models inferred by `GEMULA` fit the data better than models fitted by an existing strategy that uses MARS. We show that `GEMULA` can also be used for the analysis of mammalian gene expression data by applying it to a dataset of cultured F11 cells. This enables us to identify different sets of transcriptional regulators that are associated to early and late changes in gene expression induced by Forskolin stimulation and we gain important insights into the temporal dynamics of the regulatory network underlying neuronal outgrowth.

3.1 Introduction

The genome-wide enrichment-based analysis of gene expression using functional annotation and TFBS data in the previous chapter provides a general and flexible way of identifying relationships between TFs and target genes. Identification of such interactions is an important first step in modeling transcriptional regulatory networks. Especially in higher organisms, *combinatorial* regulation of TFs is believed to be crucial to spatio-temporal regulation of gene expression. Over the past decade, regression based statistical models have been developed that can provide a systematic way to infer plausible models of combinatorial regulation. By fitting models containing *interactions* between TFs and comparing them to simpler models without these interactions, one hopes to find evidence of condition specific regulatory interactions between TFs. Another important question, first raised by Beer *et al.* [11], that can be answered using these models is *how much* of the observed spatio-temporal variation in gene expression can be predicted or explained based on regulatory motifs occurring in gene promoters. Since not all regulatory TFBS motifs are known and because there are also other mechanisms of gene expression regulation, it is of interest to get some quantitative measure of how much of the observed variation in gene expression can be attributed to known regulatory TFBS motifs occurring in non-coding gene sequences. A statistical model where the observed gene expression is used as a response and the TFBS motifs as regressors could be used to give, at least approximately, an answer to such a question. In this chapter, we therefore consider the problem of constructing models that can be used to analyze and predict gene expression based on biologically relevant covariates.

Our main goal is to develop a pragmatic, computationally feasible (preferably fast) and biologically insightful approach to the analysis of *mammalian* gene expression data. Given some continuous response variable $Y = (Y_1, \ldots, Y_n)$ that represents observed gene expression of a set of n genes under a specific experimental condition of interest, we study models that relate a set of p covariates $X_1, \ldots, X_p$, all vectors of length n and hereafter referred to as predictors, to Y. The predictors we use will represent TFBS motifs that occur in genomic DNA regulatory sequences, but may just as well include any covariate that is *a priori* believed to be biologically related to the observed variation in Y as will become clear later on.

3.1.1 Regression approaches to modeling of gene expression data

Upon the arrival of the first high quality genome-wide gene expression studies and the availability of complete genome sequences, pioneering work on models relating DNA binding of TFs to observed gene expression was done by Bussemaker *et al.* [17]. Since at that time the DNA binding motifs of many TFs were still unknown, their work focused on methods for *ab initio* motif finding. The term *ab initio* motif finding is used to emphasize the fact that no prior knowledge of the sequences bound by TFs is used, but that such knowledge is to be derived indirectly from gene expression and DNA sequence data. In [17], this was done by constructing a motif dictionary prior to analysis. In a motif dictionary, potential TFBS DNA motifs are represented using *non-degenerate* (exact) DNA words of a predefined length L, i.e. as strings of length L composed of the 4 nucleotides `A`, `C`, `G` and `T`. The motif dictionary then consists of all unique words of length L. Predictors are constructed by counting

the number of exact occurrences of dictionary words in the regulatory DNA sequences of genes. Subsequently, the number of occurrences of a dictionary word in the regulatory DNA sequence of a gene can be used as a predictor of the gene's observed expression. Bussemaker *et al.* [17] used a linear model to asses the resulting candidate predictors in their ability to explain observed variation in gene expression. Their model assumes that the gene expression depends linearly on the number of occurrences of motifs and that the total influence of all relevant motifs is a sum over the individual ones. The implementation of the method described in [17], called `REDUCE`, was successful in identifying motifs associated to variation in gene expression during the cell cycle and sporulation in yeast. This has inspired others in attempts to build more elaborate models incorporating other mechanisms of transcriptional regulation, most notably interactions between TFs. During the past five years, there has been much interest in such models for the purpose of inferring mechanisms of transcriptional regulation and elucidating gene regulatory networks.

Several regression based tools were reviewed in Das *et al.* [25]. The methods reviewed in that paper are classified according to two criteria:

1. The type of predictors used, i.e. predictors based on a *non-degenerate* representation of motifs versus predictors based on a *degenerate* representation of motifs.

2. The type of model used to describe the relation between the response and the predictors, i.e. linear versus nonlinear.

Representations of motifs in which the likelihood or affinity of TFs that bind sequences in the DNA is determined by a model are called *degenerate*. Examples of degenerate representations include representations that use TRANSFAC PSSM models or `TRAP` models (see Section 1.1.2.2 and also [68, 85]). The representation that expresses motifs as exact DNA words, such as used for instance in [17, 121], is called *non-degenerate*. The classification in [25] makes sense, because the success of a regression based approach to the analysis of gene expression depends on the appropriateness of both the chosen type of predictors and the type of model given the available data. Non-degenerate representations have been proved to work well in yeast [17, 23, 121], where DNA regulatory regions are relatively short and well defined. Genes in higher organisms such as human and rat have a more complex structure and coding and non coding sequences are scattered across large stretches of genomic DNA. For the construction of predictors of mammalian gene expression, degenerate motif representations are deemed more appropriate [25].

In [25] the issue of which type of model, linear or nonlinear, is more appropriate for modeling mammalian gene expression data, is not addressed. Das *et al.* [23] suggest a nonlinear approach that uses Multivariate Adaptive Regression Splines (MARS). The nonparametric MARS regression methodology, which we will discuss in detail in Section 3.2.7, was introduced by Friedman [30] as a natural extension of linear models that allows nonlinearities and general interactions between regressors and incorporates model selection in a systematic way. An argument that is given in [23] to motivate the use of MARS is the following. The relationship between the binding of a TF to the DNA and the transcriptional response of a direct target gene is governed by a biophysical model. The binding affinity for a specific DNA sequence of proteins that bind the DNA depends on so-called binding free

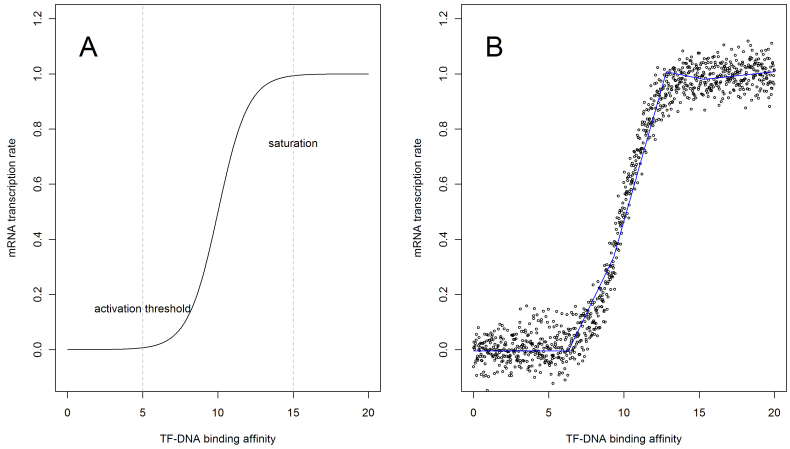

Figure 3.1: (**A**) The observed rate of transcription as a function of TF-DNA binding is believed to follow a sigmoidal pattern. (**B**) Fitted responses for a MARS model fitted using a noisy sample of size 1000 from the function in (**A**).

energy for the protein DNA interaction. Below a certain activation threshold, there is no significant transcriptional response. Just above the threshold, the gene expression response is assumed to increase nonlinearly before saturating at higher energies. A typical example of a resulting sigmoidal response curve, is shown in Figure 3.1(A). Figure 3.1(B) shows a MARS fit based on a noisy sample of size 1000. MARS uses a combination of linear splines to approximate the nonlinear sigmoidal function in Figure 3.1(A).

Although the use of MARS for regression seems to be justified from a biophysical point of view, the question remains whether regression using MARS results in better models when applied to real gene expression data. Allowing higher-order interactions between the predictors by constructing products of spline basis functions does give MARS its flexibility to model nonlinear relationships. However, it also increases the risk of overfitting. Furthermore, because MARS models are nonparametric, they require large sample sizes. A well-known pitfall common to all regression approaches is the identification of spurious regulatory interactions through models that are not appropriate for the available data or suffer from overfit. In this chapter, we will investigate the appropriateness of different regression methods and different types of predictors through a comparison based on real experimental data.

The remainder of this chapter is organized as follows. In Section 3.2 we introduce our notation and discuss concepts and methods for regression and model selection that are relevant within the context of modeling gene expression data. In Section 3.3 we focus on a linear model based regression approach and compare different methods in a simulation study. Based on the results of this study we develop an algorithm, `GEMULA`, that uses the lasso and

TRAP based predictors to model gene expression data in Section 3.4. We compare GEMULA with an existing strategy that uses MARS on yeast gene expression data in Section 3.5. We show that the use of GEMULA results in interpretable models with good fit, whereas MARS models tend to overfit the data. Finally, in Section 3.6 we apply GEMULA to a mammalian gene expression dataset to identify transcriptional regulatory interactions underlying gene expression changes in F11 cells in response to Forskolin stimulation.

3.2 Methods

In this section we discuss the problem of model selection within the context of modeling gene expression that we consider in this chapter, and introduce the methods that we use.

3.2.1 Model and notation

Suppose we observe a response vector $Y = (Y_1, \ldots, Y_n)$ that represens gene expression for a set of n genes. Additionally, let a set of p predictor variables $X_1, \ldots, X_p$ which are *potentially* biologically related to Y be given. We assume that Y and $X_1, \ldots, X_p$ are related through the following regression model

$$Y = \mathbf{X}\beta + \epsilon, \tag{3.1}$$

where

$$\mathbf{X} = [\mathbf{1} \quad f_1(X_1, \ldots, X_s) \cdots f_d(X_1, \ldots, X_s)],$$

is an unknown $n \times (d+1)$ design matrix, $\beta = (\beta_0, \ldots, \beta_d)$ is an unknown vector of regression parameters and $\epsilon = (\epsilon_1, \ldots, \epsilon_n) \sim \mathcal{N}(\mathbf{0}, \sigma^2 I_n)$. The design matrix $\mathbf{X}$ contains an intercept term and d additional predictor terms $f_k(X_1, \ldots, X_s)$, $k = 1, \ldots, d$, for some $s < p$. The predictor terms $f_k(X_1, \ldots, X_s)$ are functions of a subset of the predictor variables $X_1, \ldots, X_p$. For instance, the two main effects $f_1(X_1, \ldots, X_s) = X_2$, $f_2(X_1, \ldots, X_s) = X_5$ and the first-order interaction $f_3(X_1, \ldots, X_s) = X_2 X_5$ model the joint effect of X_2 and X_5 on Y. Furthermore, polynomial terms such as $f_4(X_1, \ldots, X_s) = X_2^2$ and $f_5(X_1, \ldots, X_s) = X_2 X_5^2$ may be considered to model curvilinear relationships.

In the following, we consider the problem of reconstructing the model in (3.1) from the observed data in the following way. Let $\mathscr{M}$ be a collection of candidate models of the form (3.1) and let $\mathscr{C}$ denote the collection of design matrices corresponding to the models in $\mathscr{M}$, i.e.

$$\mathscr{C} = \{\mathbf{X}_M : M \in \mathscr{M}\},$$

where

$$\mathbf{X}_M = [\mathbf{1} \quad f_1^M(X_1, \ldots, X_p) \cdots f_{d_M}^M(X_1, \ldots, X_p)].$$

The observed data will be used to evaluate the models in $\mathscr{M}$ through a *model selection* procedure. Model selection is the process by which a single best model or a set of models is chosen on which subsequent inference is based, often through optimization of some selection criterion over a large number of candidate models. It is a very broad topic on which a rich literature exists that deals with the model selection problem from different points of view and in different contexts (see Nishii [73], Shao [92], Burnham and Anderson [16] and Lahiri [56] and references therein). Below, we briefly discuss the underlying principles that are relevant to the regression based modeling of gene expression data that we consider in this chapter. We present the most widely used model selection criteria based on these principles and explain how these can be applied.

3.2.2 Model selection in linear models

Consider a candidate model $M \in \mathscr{M}$, with $\mathbf{X}_M \in \mathscr{C}$ and the corresponding parameter $\beta_M = (\beta_0^M, \ldots, \beta_{d_M}^M)$. When $n > (d_M + 1)$ and $\mathbf{X}_M$ is of full rank, we can use ordinary least squares (OLS) to estimate the unknown β_M by minimizing the quadratic loss

$$\hat{\beta}_M = \arg\min_{\beta} ||Y - \mathbf{X}_M \beta||^2.$$

The *goodness-of-fit* of a regression model M can be quantified in terms of the distance between Y and $\hat{Y}^M$, where $\hat{Y}^M$ denotes the vector of fitted response values under M. One widely used statistic is the coefficient of determination R^2, given by

$$R^2(M) = 1 - \frac{\sum_{i=1}^n (Y_i - \hat{Y}_i^M)^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2}, \tag{3.2}$$

where $\bar{Y}$ is the sample average of the observed responses. This statistic is often given the interpretation of the "percentage of variation explained by the model". In principle, we could use R^2 to evaluate and compare different candidate models, and select an "R^2 optimal" model

$$M_{R^2} = \arg\max_{M \in \mathscr{M}} \quad 1 - \frac{\sum_{i=1}^n (Y_i - \hat{Y}_i^M)^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2},$$

where $\hat{Y}^M = \mathbf{X}_M \hat{\beta}_M$. When one tries to reconstruct a relationship between Y and $X_1, \ldots, X_p$ from a small and noisy sample of observations, it is important the consider the following problem. It is generally unavoidable that minimization of a lack-of-fit criterion, like the quadratic loss $\|Y - \hat{Y}^M\|^2$, over a large collection $\mathscr{M}$ that contains models with different numbers of parameters, without penalization of model complexity leads to models that only "explain" variation in the particular y that was observed. In that case, the fitted model does not generalize well, i.e. it is not appropriately informing us with respect to the underlying relationship between Y and $X_1, \ldots, X_p$. This phenomenon is referred to as *overfitting*. It is due to the fact that models with a large number of predictors and many fitted parameters have many degrees of freedom. Hence, there are statistical arguments to look for models which trade-off lack-of-fit and model complexity by incorporating penalties for model complexity in model selection criteria.

The R^2 in Equation (3.2) always increases when more predictor terms are added to a candidate model M and can hence easily be artificially inflated. For that reason, it is not a useful model selection criterion. In order to find good trade-offs between model fit and model complexity when comparing many different models, various model selection criteria have been developed with a strong theoretical foundation. Among the most frequently used criteria are Akaike's information criterion (AIC) and the Bayesian information criterion (BIC).

3.2.3 Model selection criteria

Akaike [3] proposed a model selection criterion which has a rigorous foundation in information theory. He suggested the use of Kullback-Leibler information as a basis for model selection. For a candidate regression model M, the Akaike's Information Criterion (AIC) is given by

$$\mathrm{AIC}(M) = n\log(\hat{\sigma}^2_M) + 2(d_M + 1),$$

where

$$\hat{\sigma}^2_M = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i^M)^2,$$

is the estimate of the noise variance σ^2, $\hat{Y}_i^M$ are the fitted values corresponding to model M obtained using OLS and $(d_M + 1)$ is the total number of estimated parameters. Another well established criterion is the Bayesian information criterion (BIC). The *lack-of-fit* measure in AIC and BIC are the same, but BIC imposes a stronger penalty for model complexity, i.e.

$$\mathrm{BIC}(M) = n\log(\hat{\sigma}^2_M) + \log(n)(d_M + 1).$$

This difference between AIC and BIC is important, because the asymptotic properties of AIC and BIC are different. Based on the work of Shao [92], one generally distinguishes between efficient criteria, of which AIC is a well-known example and consistent criteria, of which BIC is a representative (see [73, 92, 94]). For practical regression based applications in biology and the social sciences, optimizing AIC-like criteria is believed to be good statistical practice [16]. However, there are also arguments in favor of BIC-like criteria. Hence, based on theory alone, there is no general agreement to favor a particular criterion because the theoretical properties depend on assumptions regarding the "true" underlying model (see for instance [94] and the comment by Zhang in Shao [92]). Furthermore, the AIC criterion as proposed by Akaike is not necessarily optimal for small samples. Sugiura [102] proposed a modified small sample modification of AIC that contains an additional bias correction term. This corrected AIC_c is given by

$$\mathrm{AIC}_c(M) = \mathrm{AIC} + \frac{2(d_M + 1)((d_M + 1) + 1)}{n - (d_M + 1) - 1}.$$

In [16], use of AIC_c over the original AIC is advocated if $n/(d_M + 1)$ is less than around 40.

3.2.4 Stepwise methods based on a selection criterion

Model selection based on AIC amounts to solving the following optimization problem

$$M_{\mathrm{AIC}} = \arg\min_{M \in \mathscr{M}} \mathrm{AIC}(M), \tag{3.3}$$

and similarly, for BIC

$$M_{\text{BIC}} = \arg\min_{M \in \mathscr{M}} \text{BIC}(M). \tag{3.4}$$

In applications, the number of candidate models to be considered in $\mathscr{M}$ is often large. Consequently, the optimization problems in (3.3) and (3.4) are computationally intractable and one has to resort to heuristic methods. Stepwise regression methods search through $\mathscr{M}$ by adding and/or removing predictors in a step-by-step fashion from a specified initial model in $\mathscr{M}$. At each step, all possible additions and/or deletions of a single predictor term are evaluated and the term that results in the largest change in the model selection criterion that is being optimized is added/deleted, until a (local) optimal value is found or another stopping criterion is met.

3.2.5 Penalized least squares and the lasso

A recent trend in statistics which is becoming increasingly popular also in practical applications is the use of regularized regression approaches such as ridge regression and the lasso. The lasso was introduced by Tibshirani [105]. For a given matrix $\mathbf{X}_M$ containing candidate predictors, let

$$T(\beta_M) = \sum_{j=1}^{d_M} |\beta_j^M|.$$

Lasso estimates of β_M are solutions of the minimization problem

$$\min_{\beta_M} \sum_{i=1}^{n} \left(Y_i - \beta_0^M - \sum_{j=1}^{d_M} Z_{ij} \beta_j^M \right)^2 \quad \text{subject to} \quad T(\beta_M) \leq t, t \in \mathbb{R}^+, \tag{3.5}$$

where $Z_j = f_j^M(X_1, \ldots, X_p)$. As is clear from (3.5), lasso solutions depend on a shrinkage parameter t. Because of the particular geometry of the minimization problem, i.e. due to the L_1 constraint in (3.5), decreasing t results in coordinates $\hat{\beta}_j^M$ in the obtained lasso solution becoming zero exactly. Hence, by varying t the lasso is effectively performing shrinkage and variable selection simultaneously. Good values of t are typically found using cross-validation or by minimizing criteria such as AIC or BIC. In [26], an efficient algorithm is described to compute the entire path of all lasso solutions. An attractive feature of the lasso is that the computational complexity of the algorithm described in [26] is of the same order as the algorithm used to obtain the OLS solution.

3.2.6 Random forests

The random forest algorithm, developed by Breiman [14, 13], is an ensemble classifier that has its roots in machine learning. Although developed within the classification framework, the trees *grown* by the random forest algorithm can also be used for regression. Methods using tree-based regression to identify TFBS motifs underlying variation in gene expression in yeast have been described by Phuong *et al.* [77] and Xiao and Segal [119]. Suppose we

observe p predictors $X_1, \ldots, X_p$ and a response $Y = (Y_1, \ldots, Y_n)$, all vectors of length n. In a regression tree, nodes represent subsets of observations. The root contains all n observations. Each node specifies a binary partition of the observations into 2 descendant nodes based on a split, which is a function of the predictors. A typical example of a split function for a node A is "$X_j \geq z$", for some number z. This means that A is split into $B = \{i \in A : x_{ji} \geq z\}$ and $C = \{i \in A : x_{ji} < z\}$.

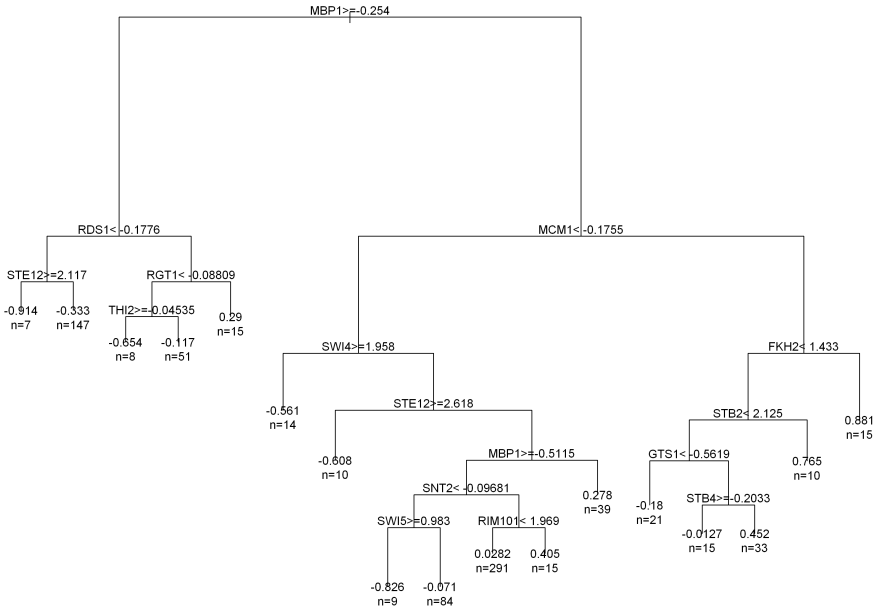

Figure 3.2: Example of a regression tree fitted on yeast cell cycle gene expression data. For a detailed description of this dataset, see Section 3.3.1.

Figure 3.2 contains an example of a regression tree constructed for yeast cell cycle data. In this dataset the response Y is the observed gene expression of 790 cell cycle regulated yeast genes at 56 minutes following α-factor synchronization. The predictors consist of variables that represent binding affinity of TFs for binding to the promoter regions of the cell cycle genes. The terminal nodes define a partition of all 790 observations. Fitted values are determined by averaging over all observations in each terminal node. For instance, this dataset contains 7 genes for which both $\texttt{MBP1} \geq -0.254$ *and* $\texttt{STE12} \geq 2.117$ hold, and the average response of these genes is -0.914, see Figure 3.2. The crux in constructing good regression trees lies in determining good splits and the decision when to stop splitting, see Breiman *et al.* [14].

In standard regression tree modeling, at each node the best split is chosen by maximizing

some split loss-criterion over a large number of candidate splits involving *all* predictors. Because this procedure is prone to overfitting, regression trees in a random forest use subsets of predictors randomly chosen at each node to determine the split. The random forest algorithm combines a large number (ensemble) of different regression trees and uses machine learning strategies to integrate the individual trees into a *forest* regression model. One prominent feature is the use of bootstrap aggregating (or *bagging*), i.e. the regression trees in the ensemble are grown using independent bootstrap samples and prediction is determined by majority voting in the ensemble.

3.2.7 MARS

Friedman [30] introduced a nonparametric regression methodology that uses multivariate adaptive regression splines (MARS). MARS can be viewed as a natural extension of linear regression models that allows for modeling of nonlinearities and general interactions between variables. An essential part of the MARS methodology are the so-called *hinge* functions that are used, typically in pairs. Given any predictor X_j and a real number k_j called the *knot*, MARS builds regression models using expansions in piecewise linear spline functions of the following form

$$(X_j - k_j)_+ = \begin{cases} X_j - k_j, & \text{if } X_j > k_j, \\ 0, & \text{otherwise,} \end{cases}$$

and

$$(k_j - X_j)_+ = \begin{cases} k_j - X_j, & \text{if } X_j < k_j, \\ 0, & \text{otherwise.} \end{cases}$$

For each predictor variable X_j, MARS considers such functions, with the knot values k_j ranging over the observed values x_j to give a collection B_j of basis functions

$$B_j = \{(X_j - k_j)_+, (k_j - X_j)_+ : j = 1, \ldots, p \qquad k_j \in \{x_{j1}, \ldots, x_{jn}\}\},$$

Figure 3.3, shows an example of a pair of mirrored hinge functions, for a predictor variable X_1 and a knot at the value 2.134.
The basis functions for the individual predictors together form an entire collection $\mathscr{C} = \cup_j B_j$. The MARS model has the following form

$$Y = \beta_0 + \sum_{l=1}^{L} \beta_l b_l(X_1, \ldots, X_p) + \epsilon,$$

where $b_l(X_1, \ldots, X_p)$ are elements of $\mathscr{C}$ or products of two or more of such functions and $\epsilon = (\epsilon_1, \ldots, \epsilon_n)$ is a vector of i.i.d. errors with $\mathbb{E}(\epsilon_i) = 0$. MARS uses products of elements of $\mathscr{C}$ to model interactions up to a certain degree that is controlled by a user defined *degree* parameter κ. When $\kappa = 1$, MARS builds additive models without interactions, whereas first-order and second-order interactions can be modeled using $\kappa = 2$ and $\kappa = 3$ respectively. The model selection procedure implemented in the MARS algorithm consists of two phases. In the forward phase, starting from an initial model containing an intercept term only, terms

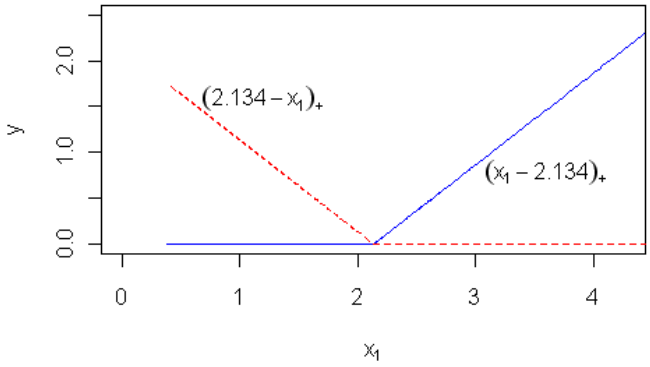

Figure 3.3: Example of a pair of basis functions used by MARS

are sequentially added using a greedy strategy until a maximum number of model terms is reached. In a successive pruning phase, terms are removed from the model in order to find an optimal model according to a modified *generalized cross-validation* (GCV) criterion. Let $\hat{Y}^M$ denote the vector of fitted response values. Then, for a MARS model M with $L+1$ model terms (including an intercept), the GCV is given by

$$\mathrm{GCV}(M) = \frac{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i^M)^2}{(1 - \frac{\mathrm{Q}(M)}{n})^2},$$

where

$$\mathrm{Q}(M) = (L+1) + \lambda k_M$$

is a function of the number of estimated parameters $(L+1)$ and a penalty term λk_M, where k_M is the number of knots in the model and λ is a chosen constant. Hence, λ can be used to penalize for the inclusion of knots in the model. Friedman suggests values of λ between 2 and 4 for use in practical applications [30]. In order to assess the goodness-of-fit of MARS models fitted using their `MARSMOTIF` algorithm, Das *et al.* [23, 24] use a R^2-like statistic $\Delta\chi^2$ which is given by

$$\Delta\chi^2(M) = 1 - \frac{\sum_{i=1}^{n}(R_i - \bar{R})^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}, \tag{3.6}$$

where $R_i = Y_i - \hat{Y}_i^M$, $\bar{R} = \frac{1}{n}\sum_{i=1}^{n} R_i$ and $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$.

3.3 Simulation study

In this section, we conduct a simulation study in order to compare different model selection methods. We first derive a model, hereafter called the *pilot model*, that is appropriate within the context of the application that we are interested in, i.e. one that is derived from real experimental gene expression data and that can be viewed as an approximation of the "true" data generating model. Next, we simulate artificial gene expression datasets from this pilot model and compare fitted models obtained by applying different model selection methods to the simulated data. In this way, we can quantify the tendency of the methods to overfit the observed data. An additional advantage of a simulation study is that we can vary the sample size and the variance of the noise in the data and study the possible effects they may have on the outcome.

As candidate "true" data generating models we consider linear models, because it has been shown that such models are a practically useful approximation [17, 121]. Hence, it is of practical relevance to know which model selection procedure would give us the model with the best fit if the data *were* generated by such a model. We fit a linear model with a stepwise variable selection algorithm to real yeast gene expression data and the resulting model will be our pilot model from which the artificial gene expression data will be simulated.

3.3.1 The pilot model

Gene expression time-course profiles of synchronized yeast cultures progressing through the different stages of the cell cycle were measured by Spellman *et al.* [96]. In the cell cycle study performed by Spellman *et al.* three independent sets of experiments were done using different methods to synchronize the yeast cells. Here, we consider the α-factor arrest experiments, which contain measurements of all yeast genes at 18 different time points following synchronization, spanning three complete cell cycles (periods). In [96], 800 yeast genes were identified as being cell cycle regulated. The expression profiles of these 800 genes display a clearly distinguishable periodic pattern, which is well known to be governed by a number of different transcription factors. In fact, some of the TFs are themselves "periodically" expressed, although this is not a necessary condition for a TF to regulate cell cycle periodic genes.

We consider the expression of the 800 cell cycle regulated genes at 56 minutes following α-factor arrest. The microarrays used to measure cell cycle expression are two-channel arrays, hence the observed gene expression values correspond to log-ratios of expression at 56 minutes compared to control. The 56 minute timepoint in the Spellman *et al.* cell cycle data corresponds roughly to the cell cycle phase just after the transition from `G2` to `M`. The motivation for this particular time-point is that the transition from `G2` to `M` is known to be coordinated transcriptionally. The Spellman gene expression data contain some missing data, resulting from spots on the microarray for which no accurate log-fold expression ratios could be obtained, but given the high degree of correlation between periodically co-expressed transcripts, the missing values can be estimated in a reliable way. We use the `KNNImpute` algorithm developed by Troyanskaya *et al.* [106] to impute missing gene expression values.

From the experimentally derived DNA binding sites published by MacIsaac *et al.*[66], we

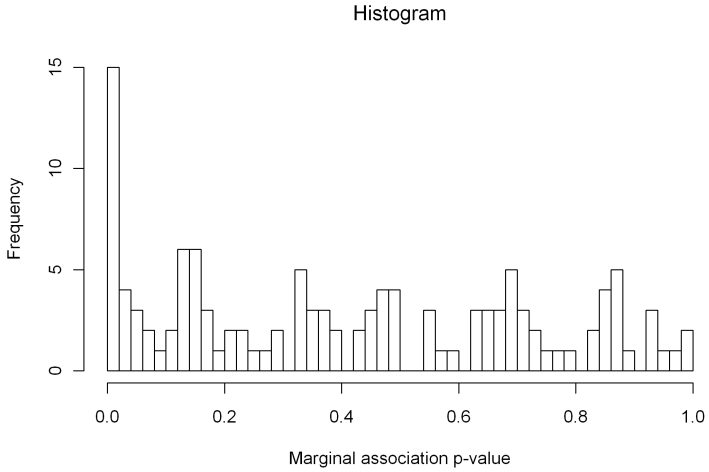

Figure 3.4: Marginal association p-values of all 123 different TRAP predictors.

extract 123 different position frequency matrices representing models of the DNA sequences bound by the different transcription factor proteins. We use the `TRAP` [85] tool to calculate DNA binding affinities for binding to the genomic DNA sequences from 1 bp to 1000 bp directly upstream of the cell cycle regulated genes. The genomic sequences were obtained from SGD [18]. For 10 out of the 800 cell cycle genes, we could not match the IDs to IDs of the genomic sequences from SGD, which brings the total sample down to 790. The resulting predictors $X_1, \ldots, X_{123}$ thus represent binding affinities of 123 yeast DNA binding TFs.

In order to fit the pilot model, we first need to define a set of appropriate linear candidate models $\mathscr{M}$ to consider (see Section 3.2.4). It is known that cooperation between TFs is important for cell cycle gene regulation and others have reported pairwise interactions between yeast cell cycle TFs [121, 23]. We therefore focus on identifying main effects and effects corresponding to interaction effects between the predictors. With 123 variables, there are 7503 possible candidate pairwise effects to consider. Since we expect only a subset of predictors to be truly associated to Y, we do a univariate-screening to select the predictors most strongly associated to Y univariately. We allow only interaction terms between these predictors in candidate models. For all candidate predictors X_j, for $j = 1, \ldots, 123$, we calculate t-statistics indicating the significance of the estimate of the regression coefficient β_j in the model $Y = \beta_0 + \beta_j X_j + \epsilon$. The p-values corresponding to the tests $H_{0j} : \beta_j = 0$ can be used to rank the predictors. Figure 3.4 contains a histogram plot of all 123 marginal p-values. The lowest observed p-value (unadjusted for 123 tests) is 2.1×10^{-18} which indicates strong evidence of association. There are 21 predictors with a marginal unadjusted p-value smaller than 0.05. Based on these results, we fit our pilot model by limiting the set of candidate predictor terms to

1. main effects for all predictor variables X_j, $j = 1, \ldots, 123$,
2. pairwise interaction effects for the 25 most strongly univariately associated predictor variables.

This brings the total number of candidate terms to $123 + 300 = 423$. We use the `step()` function in `R` [79] to perform model selection. This function implements a greedy stepwise search strategy that considers both forward and backward moves, starting from an initial model that contains an intercept term only. We use the AIC to evaluate and compare visited models in the stepwise search. Our resulting pilot model contains 57 terms, 33 main effects and 23 first-order (pair-wise) interactions and an intercept term. The observed multiple (unadjusted) R^2 for the pilot model is 0.41. The estimate of the variance of the noise is $\hat{\sigma}^2 = 0.15$.

3.3.2 Model selection on simulated data

We compare the performance of different model selection methods on simulated data at three different noise levels, referred to as *low*, *medium* and *high* respectively. From previous studies [17, 23, 121] and our own data, which we analyze in Section 3.6, we conclude that in practice we may expect the number of genes, i.e. the sample size n, to approximately range between 500 and 2000. Therefore, at each noise level, we run two series of independent simulations, one with sample size $n = 790$ and one with $n = 2000$. In order to simulate data, we use the design matrix and the estimated vector of regression coefficients of the pilot model. When $n = 2000$, we add rows to the design matrix by sampling genes uniformly at random from $6717 - 790 = 5927$ not cell cycle regulated yeast genes for which we have data available. In Table 3.1, we give an overview of properties of the data generating model that we use in our simulation study. The candidate models in $\mathscr{C}$ that we allow the methods to consider contain the following terms.

1. All 123 main effects of the predictors $X_1, \ldots, X_{123}$.
2. All 23 first-order interactions present in the pilot model.
3. 154 additional first-order interactions not present in the pilot model.

The methods we include in our comparison are listed in Table 3.2. Because the lasso produces a whole path of solutions indexed by a shrinkage parameter, lasso models depend on the chosen value for that parameter. We select the optimal shrinkage parameter as the optimizer of a model selection criterion and compare the AIC, the finite sample corrected version of AIC (AIC_c) and BIC. For each round of simulations, i.e. for a fixed sample size n and fixed noise level, results are based on 200 independent realizations of Y from the data generating model. All model selection methods from Table 3.2 are applied 200 times to fit a model for Y using $X_1, \ldots, X_{123}$ as predictors. We compare the performance of all different methods based on the root mean squared prediction error (RMSE). In each run k, each method fits a regression model using $X_1, \ldots, X_{123}$ and y_k. For method j in run k, this fitted model defines a rule $\hat{f}_{jk}$ that maps values of the predictors $X_{1i}, \ldots, X_{pi}$ for any given gene i to some fitted

Property description	Values used in simulations
sample size n (number of genes)	$\{790, 2000\}$
noise variance σ^2	$\{0.87, 0.15, 0.065\}$
number of predictors in the pilot model	25
number of terms in the pilot model	56
number of candidate predictors considered	123
number of candidate terms considered	300

Table 3.1: Properties of the pilot model and the candidate models in the simulation study.

response value $\hat{y}_{ijk}$, i.e. $\hat{y}_{ijk} = \hat{f}_{jk}(X_{1i}, \ldots, X_{pi})$. Let $\mu = \mathbf{X}_p \beta_p$, where $\mathbf{X}_p$ is the design matrix of the pilot model and β_p the vector of regression coefficients of the pilot model. The RMSE of $\hat{f}_{jk}$ is then given by

$$\text{RMSE}(\hat{f}_{jk}) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\mu_i - \hat{y}_{ijk})^2}.$$

Shorthand	Description
ols	OLS multiple linear regression using all candidate terms
step.aic	OLS with stepwise selection based on AIC
step.bic	OLS with stepwise selection based on BIC
lasso.aic	L_1-penalized lasso regression based on AIC
lasso.aic.c	L_1-penalized lasso regression based on AIC_c
lasso.bic	L_1-penalized lasso regression based on BIC
rf	Random forests regression

Table 3.2: Ensemble of model selection methods used in the simulation study.

3.3.3 Results

We present the results of the simulations in box-and-whisker plots that contain RMSEs for the different methods from Table 3.2. Figure 3.5, Figure 3.7 and Figure 3.8 show the results for sample size $n = 790$ as in our pilot study of yeast data. Overall, we conclude that the lasso combined with selection of the shrinkage parameter based on AIC/AIC_c outperforms the other methods considered. The results indicate that model selection based on AIC results in lower RMSEs on average than when BIC is used for both the stepwise OLS method and

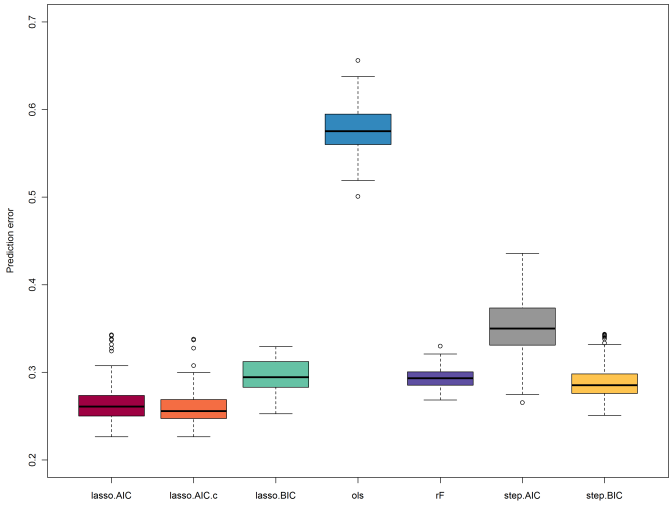

Figure 3.5: Box plots of RMSEs of different model selection methods obtained on simulated data with $n = 790$ and a high noise level.

the lasso. Only when the noise level is high and the sample size is small, stepwise BIC tends to do better than stepwise AIC (see Figure 3.5). Another remarkable result is that regression using random forests does well when the noise variance is high, but that its performance drops radically when noise levels become smaller. Figure 3.6 contains results obtained with a larger sample size of $n = 2000$, but with the same noise variance as in Figure 3.5. The results in these two figures are similar. In fact, for the other two noise levels we did not find great differences when the sample size was increased from 790 to 2000 and we present only the results for $n = 790$. Figure 3.8 represents an optimistic scenario in which the noise variance is lower than we may expect in practice. In that case, the performance of stepwise selection with AIC gets closer to the performance of the lasso, although results obtained with stepwise selection are much more variable. Moreover, stepwise selection methods are computationally much more demanding than lasso regression, which becomes more significant as the number of candidate models to be considered increases. From these results we conclude that a regression approach using the lasso compares favorably to regression using random forests and OLS stepwise model selection procedures on our simulated expression data. The lasso is relatively fast, can handle a large number of candidate predictors and is capable of selecting a parsimonious regression model. Therefore, in the next section we develop an approach that uses the lasso to analyze gene expression data.

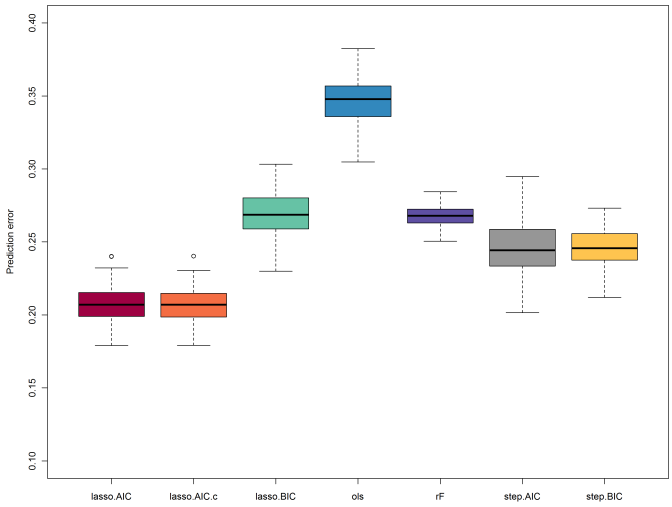

Figure 3.6: Box plots of RMSEs of different model selection methods obtained on simulated data with $n = 2000$ and a high noise level.

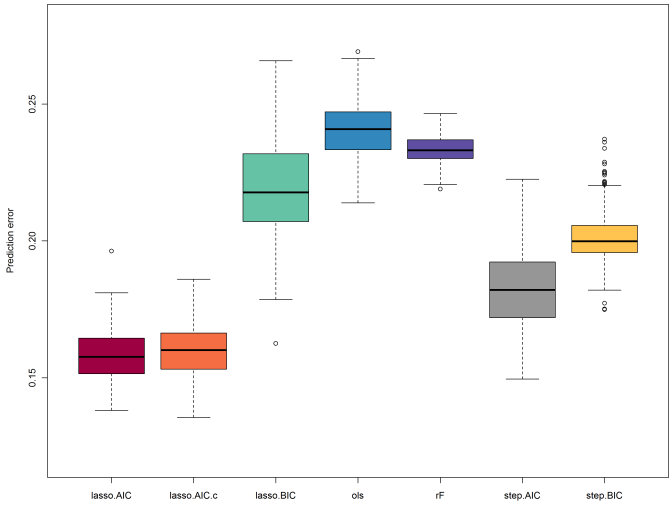

Figure 3.7: Box plots of RMSEs of different model selection methods obtained on simulated data from model with $n = 790$ and a medium noise level.

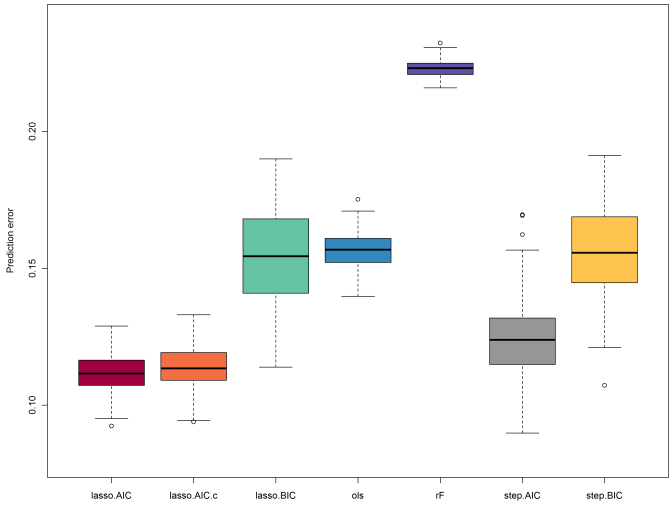

Figure 3.8: Box plots of RMSEs of different model selection methods obtained on simulated data with $n = 790$ and a low noise level.

3.4 GEMULA: gene expression modeling using lasso

In this section we introduce GEMULA, a three-stage procedure based on the lasso to model gene expression data that uses biologically relevant variables as predictors. Let a response variable $Y = (Y_1, \ldots, Y_n)$ that represents gene expression for a set of n genes and additionally a set of p predictor variables $X_1, \ldots, X_p$, all vectors of length n, be given. Our aim is to provide a generally applicable algorithm that is reasonably fast, that considers a wide enough range of plausible models and that is capable of selecting a good regression model for the data that represents a sensible trade-off between bias and variance. The approach will consist of the following three stages.

I **Predictor order determination**. Determination of the order in which predictors are to be considered in the model building stage. Order determination is based on a lasso fit considering all main terms of the predictor variables.

II **Model building**. Generation of a small number of candidate models. Each of the models is selected from a different large subset of candidate models that is determined by *a priori* chosen parameters such as maximum number of allowed predictor terms, maximum order of interaction and inclusion of nonlinear terms. Within each subspace a representative model is identified using lasso-AIC shrinkage and selection.

III **Model selection/validation**. Selection of the best model among candidate models generated in the model building stage, through cross-validation.

To describe the algorithm we first introduce some notation. We use the same notation as in Section 3.2.1 to describe candidate regression models M and the corresponding matrices $\mathbf{X}_M$. Recall (see Section 3.2.5) that for $t \in \mathbb{R}^+$, the lasso estimate of β_M is determined by

$$\min_{\beta_M} \sum_{i=1}^{n} \left(Y_i - \beta_{M0} - \sum_{j=1}^{d_M} \beta_{Mj} Z_{ij} \right)^2 \quad \text{subject to} \quad \sum_{j=1}^{d_M} |\beta_{Mj}| \leq t, \tag{3.7}$$

where $\beta_M = (\beta_{M0}, \ldots, \beta_{Md_M})$ and $Z_j = f_j^M(X_1, \ldots, X_p)$. We use the `lars` algorithm [27] repeatedly to solve the different penalized least squares regression problems. The parameter $t \in \mathbb{R}^+$ is used to index the entire lasso path. The algorithm proceeds in steps, indexed by k, for $k = 0, \ldots, K$, and we identify the lasso solution at step k by $\hat{S}_M(t_k)$. We denote the entire path by $\mathscr{S}_M = \{\hat{S}_M(t) : t \in \mathbb{R}^+\}$. We let $\hat{\beta}_M^k = (\hat{\beta}_{M0}^k, \ldots, \hat{\beta}_{Md_M}^k)$ denote the estimate of β_M corresponding to the lasso solution at step k, $\hat{\mu}_M^k = \mathbf{X}_M \hat{\beta}_M^k$, $df(\hat{\mu}_M^k)$ the corresponding degrees of freedom, $\mathscr{B}_M^k = \{j : \beta_{Mj}^k \neq 0\}$, and $b_M^k = |\mathscr{B}_M^k|$. Initially, when $k = 0$, $\mathscr{B}_M^k = \emptyset$ and $b_M^k = 0$.

Zhou *et al.* [123] show that model selection criteria (see Section 3.2.3) can also be used in lasso model selection. To select a model along the path $\mathscr{S}_M$, GEMULA optionally uses either BIC, AIC or AIC_c. Motivated by the results from the simulation study, we use the AIC_c criterion when we apply GEMULA to analyze real gene expression data. Let $\text{AIC}_c(\hat{S}_M(t_k))$

denote this criterion, for $\hat{S}_M(t_k) \in \mathscr{S}_M$. Then

$$\mathrm{AIC}_c(\hat{S}_M(t_k)) = \frac{\|Y - \hat{\mu}_M^k\|^2}{n\sigma^2} + \frac{2}{n} df(\hat{\mu}_M^k) + \frac{2df(\hat{\mu}_M^k)(df(\hat{\mu}_M^k)+1)}{n - df(\hat{\mu}_M^k) - 1}. \tag{3.8}$$

It is shown in [123] that the optimal model in $\mathscr{S}_M$ according to the selection criterion can be found by minimizing (3.8) over all t_k, $k = 0, \ldots, K$ and therefore we let

$$k_M^{\mathrm{AIC}_c} = \arg\min_{t_k} \mathrm{AIC}_c(\hat{S}_M(t_k)).$$

Now we describe the three stages of GEMULA in more detail.

I In this stage GEMULA determines the order in which the input predictors may enter the candidate models. Let M_0 represent the model for which the design matrix satisfies $\mathbf{X}_{M_0} = [\mathbf{1} \quad X_1 \cdots X_p]$. Since at each step k, the index of exactly one predictor enters the set $\mathscr{B}_{M_0}^k$, GEMULA uses the mapping

$$r(j) = \min \quad \{k : j \in \mathscr{B}_{M_0}^k\}, \quad j \in \{1, \ldots, p\},$$

and its inverse r^{-1} defined by

$$r^{-1}(s) = j \qquad \Leftrightarrow r(j) = s \qquad j \in \{1, \ldots, p\}, s \in \{1, \ldots, K\}$$

to define the order.

II In this stage GEMULA generates candidate models confined to Q different candidate model subspaces. The different model subspaces are identified by three-dimensional parameters $\gamma_q = (\gamma_{q1}, \gamma_{q2}, \gamma_{q3})$, for $q = 1, \ldots, Q$, where γ_{q1} represents the maximum allowed order of interactions between terms, γ_{q2} the maximum power to which candidate predictors are raised in candidate terms and γ_{q3} represents the maximum number of candidate terms allowed in the model. The complete collection of models that are considered by GEMULA is $\mathscr{M} = \mathscr{M}_{\gamma_1} \cup \cdots \cup \mathscr{M}_{\gamma_Q}$. For the model subspace $\mathscr{M}_{\gamma_q}$ defined by γ_q, $\mathbf{X}_{\gamma_q}$ denotes the design matrix of the model in $\mathscr{M}_{\gamma_q}$ with the largest possible number of predictors confined by the order determined in step I. For instance, we can restrict GEMULA to models containing only main effects of the first 50 predictors $X_{r^{-1}(1)}, \ldots, X_{r^{-1}(50)}$ by setting $\gamma_1 = (1, 1, 50)$. We write $\mathbf{X}_{\gamma_1} = [\mathbf{1} \quad X_{r^{-1}(1)} \cdots X_{r^{-1}(50)}]$. When interactions between predictors are considered, the restrictions on the maximum number of allowed terms imposed by γ_{q3} force GEMULA to limit the number of predictors in the following way. In a model with s predictors, there are s main effect terms and $s(s-1)/2$ possible pairwise interactions. Suppose we set $\gamma_2 = (2, 1, 150)$, then GEMULA first determines

$$s^* = \max\{s \in \{1, \ldots, p\} : s + s(s-1)/2 \leq 150\},$$

and then $\mathbf{X}_{\gamma_2}$ denotes the design matrix that contains all main effects and possible interactions between the predictors $X_{r^{-1}(1)}, \ldots, X_{r^{-1}(s^*)}$. For each matrix $\mathbf{X}_{\gamma_q}$, we fit the

entire path of lasso solutions $\mathscr{S}_{\gamma_q}$ and select the optimal lasso-AIC shrinkage parameter $k_{\gamma_q} = k_{\gamma_q}^{\mathrm{AIC}_c}$. We denote the selected candidate model, i.e. the selected subset of model terms identified by $\mathscr{B}_{\gamma_q}^{k_{\gamma_q}}$, by M_q and the corresponding fitted response values by $\hat{Y}^{M_q} = \hat{\mu}_{\gamma_q}^{k_{\gamma_q}}$.

III `GEMULA` uses cross-validation to evaluate the fit of the Q candidate models. As goodness-of-fit measure, we use the R^2 statistic, because it has an intuitive interpretation that is of interest also biologically. Recall that for a candidate model M_q and corresponding fitted response values $\hat{Y}^{M_q}$, the R^2 is given by

$$R^2(M_q) = 1 - \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}^{M_q})^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}.$$

3.5 Validation on yeast data

In order to investigate whether GEMULA in combination with TRAP based predictors is capable of identifying TFBS motifs and interactions between TFs that are associated to variation in observed gene expression, we first apply it to data from yeast. Because many transcriptional regulatory interactions in yeast are known and have been experimentally verified, we gain insight into GEMULA's competitive performance through a comparison of GEMULA and MARS using different sets of predictors. Next, we show the potential of the use of GEMULA in combination with additional biologically important predictors derived from experimental data on yeast heat shock expression data.

3.5.1 Yeast cell cycle

We applied both GEMULA and MARS to analyze the yeast cell cycle gene expression data introduced in Section 3.3.1. As response variables Y we considered the observed gene expression at both 49 and 56 minutes following α-factor synchronization. To construct our TRAP predictors, we used two sets of PSSMs (see Section 1.1.2.2). The first is the set of 123 PSSMs extracted from the ChIP-chip data published by MacIscaac *et al.*[66], for which the corresponding DNA binding TFs are known. We already introduced these in Section 3.3.1 and refer to the derived TRAP predictors as TRAP MRM. The second is deduced from a set of 666 PFMs that were obtained from data published by Beer and Tavazoie [11]. The resulting predictors are referred to as TRAP 666. Beer and Tavazoie used AlignACE to predict TFBS motifs *ab initio* from yeast gene expression data measured under 255 different experimental conditions, including many environmental stress conditions from data published by Gasch *et al.* [35] and the same cell cycle data from Spellman *et al.* that we analyze here. They obtained 615 *ab initio* predicted motifs and complemented their set of motifs with 51 known motifs. Hence, to a great extent these 666 PFMs correspond to motifs for which the DNA binding TF was unknown at the time. Note that the set of 615 "new" motifs identified in [11] may be redundant and contain variants of motifs recognized by a single TF. This complicates interpretation of the results obtained using the TRAP 666 predictors. However, we included this dataset because it resembles a more "complete" set of motifs which we can compare to an exhaustive set of non-degenerate DNA words. Therefore, we also included a set of non-degenerate predictors. We used a motif dictionary of all possible DNA words of length 6 compiled by Zhang *et al.* [121]. The predictors are constructed as counts of the occurrence of each dictionary word in the DNA regulatory sequences of yeast cell cycle regulated genes. We refer to these predictors as 'Dictionary'.

To generate and fit candidate models with GEMULA, we need to specify the three-dimensional γ tuning parameter. For the yeast cell cycle data, chose $Q = 4$ and used $\gamma_1 = (1,1,600)$, $\gamma_2 = (2,1,600)$, $\gamma_3 = (3,1,600)$, $\gamma_4 = (2,2,600)$ and we refer to the models selected by GEMULA as M1, M2, M3 and M4 respectively. We used the R package `earth` which provides an implementation of MARS as described in the original paper of Friedman [30] for the analysis. Das *et al.* [23] performed a candidate reduction step based on the Kolmogorov-Smirnov test to reduce the number of candidates considered by their MARSMOTIF algorithm. We noted gains in performance in terms of both model fit and run-time when the number of

candidates is reduced prior to running the MARS algorithm. Therefore, for a given response Y, we selected the 50 predictors that univariately are most strongly associated to Y as input to MARS, using the same univariate screening we applied prior to fitting our pilot model in Section 3.3.1. For MARS models, the allowed order of interactions between predictors can be controlled using the `degree` parameter κ (see Section 3.2.7). We used the values $\kappa = 1$ to fit additive MARS models and $\kappa = 2$ to include first order interactions between predictors. Another parameter in MARS which is important for model selection is the parameter λ which controls the penalization for the inclusion of knots in the fitted MARS model. We ran MARS and report results for values within the range $2 \leq \lambda \leq 4$ suggested in [30] and in order to be conservative also include results obtained with higher, more stringent values.

The results are presented in Tables 3.3 and 3.4. In these tables, the column `Model type` contains information about the set of candidate models that were considered. The columns `Model P` and `Model T` contain the number of predictors and the number of terms in the fitted model, respectively. We conclude that the motif dictionary used by Zhang *et al.* is the "best" set of predictors for yeast cell cycle gene expression in terms of $\bar{R}^2_{cv}$. Note the high number of main terms selected by `GEMULA` when counts are used. Since TFs recognize degenerate motifs, the Dictionary is very redundant and the 199 "different" motifs selected by `GEMULA` will correspond to a much lower number of actual motifs. We should also keep in mind that the Dictionary predictors are discrete counts whereas our `GEMULA` procedure assumes real-valued predictor variables. For modeling of the main effects, `GEMULA` can be used with discrete predictors and this actually leads to models with high $\bar{R}^2_{cv}$s. Interactions between the predictors, however, may not be appropriately modeled using simple cross-product terms when the predictors are discrete. We clearly see the benefit of modeling interactions between predictors with `GEMULA` when the `TRAP` MRM predictors are used. For both time-points, `GEMULA` models that include interaction terms between `TRAP` MRM predictors consistently outperform `GEMULA` models that exclusively contain linear main effects. Also note that the $\bar{R}^2_{cv}$s obtained on the 56 minute time-point are higher than those obtained on the 49 minute time-point.

The results of the analysis using MARS on the same yeast cell cycle data are shown in Table 3.4. We find that models inferred using MARS perform poorly in terms of $\bar{R}^2_{cv}$. With the degree parameter set to 2, we find negative $\bar{R}^2_{cv}$ values, which is indicative of overfit. When the degree parameter is set to 1, we do find positive $\bar{R}^2_{cv}$ values, but much lower than the corresponding $\bar{R}^2_{cv}$ for models fitted using `GEMULA`. We note that the $\Delta\chi^2$s of the MARS models we fit are in the same range as reported by Das *et al.* in [23]. For instance, for the MARS model fitted on the 49 minute time-point using the Dictionary predictors, we find $\Delta\chi^2 = 0.30$ for the model with $\kappa = 1$ and $\lambda = 4$ and $\Delta\chi^2 = 0.26$ for the model with $\kappa = 2$ and $\lambda = 10$. We compare this to the reported $\Delta\chi^2 = 0.26$ in [23] for a MARS model fitted on the 49 minute time-point, also using a predictor set representing discrete counts of words.

Let us take a closer look at the predictors that underly the variation in observed yeast cell cycle gene expression according to `GEMULA`. Here, we only consider the identity of the predictors that occur in main terms and interactions in models fitted by `GEMULA`. We postpone a comprehensive analysis of the relative *importance* of the different predictors to Chapter 4. For ease of interpretation, we focus on the `GEMULA` models that use the `TRAP MRM` predictors. The `M2` models at the time-points 49 and 56 contain the well known

cell cycle TFs MBP1, MCM1, SWI5, SWI6, FKH2, STE12, ACE2 and DIG1 among others. For all these, there is strong evidence in the literature that they are involved in the regulation of cell cycle genes, see for instance [107]. Zhang *et al.* report a model containing 35 terms in their analysis of the cell cycle gene expression data [121]. The DNA words of the predictors in their inferred model also include words that represent binding motifs for MBP1, FKH1/2, FKH2, SWI4, SWI5, SWI6 and MCM1. Both the GEMULA models that we infer and the models inferred by Zhang *et al.* contain pairwise interactions between predictors. In general, pairwise interactions are difficult to validate experimentally. For yeast cell cycle gene expression, there is strong evidence in the literature for cooperation of certain pairs of TFs. The interacting pairs identified by the M2 model include MCM1:FKH2, SWI5:ACE2 and SWI5:FKH2. Interactions between these pairs of TFs are all experimentally verified (Tsai *et al.*[107], Banerjee and Zhang [9]).

3.5.2 Yeast heat shock

The 615 "new" TFBS motifs identified in [11], which are part of our TRAP 666 set, were based on a clustering of gene expression data that, apart from the cell cycle data from Spellman *et al.*, included gene expression in response to changes in environmental conditions. As environmental stress is known to trigger strong transcriptional responses, we further validate GEMULA on gene expression data measured in yeast cultures exposed to a sudden change in temperature. Gasch *et al.*[35] published genomic expression patterns of the yeast *Saccharomyces cerevisiae* in cells that were exposed to various changes in environmental conditions including heat shock, hypo-osmotic shock and amino acid starvation. Their most important finding was that a large set of genes showed a comparable drastic transcriptional response to almost all examined stress conditions. They coined the term 'environmental stress response' (ESR) to describe this phenomenon. The group of approximately 900 ESR-induced genes can be further divided into two subsets based on the direction of the induced expression changes with respect to baseline conditions, i.e. into sets of consistently up-regulated and down-regulated genes. The heat shock expression dataset contains time-course profiles of the ESR genes at 8 different time-points, ranging from 5 to 80 minutes post-shock. Here, we consider the observed gene expression at 20 minutes as our response variable Y. Analysis of the other time-points yielded similar results.

Additionally, we address the following issue. Observed variation in rates of transcription can be influenced by other factors besides binding of TFs. Therefore it is of interest to know if models fitted using predictors representing TFBS motifs exclusively, can be further improved through inclusion of additional relevant predictors. Examples of such predictors are experimentally determined data on nucleosome occupancy and TF binding data obtained by ChIP-chip assays. During the past few years high resolution maps have been published on nucleosome occupancy in yeast. Eukaryotic genomic DNA is tightly packaged in a structure called chromatin, of which nucleosomes are the fundamental repeating unit. This tight packing of the DNA is necessary to ensure that the long linear DNA molecules fit inside the nucleus. The chromatin structure accomplishes this while still providing accessibility to the DNA for active transcription. Nucleosomes are critical to the organization and maintenance of chromatin and the positioning and modification state of nucleosomes are known to

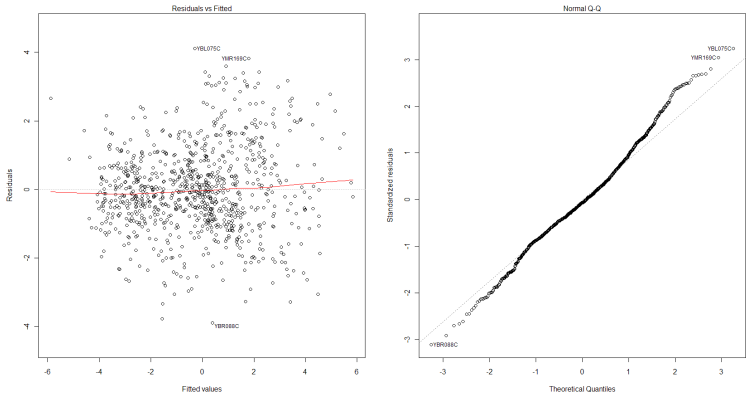

Figure 3.9: Evaluation of the GEMULA M2 model fitted using TRAP 666 + NUC predictors on yeast heat shock gene expression data. The plot on the left shows the model residuals against the fitted values. The plot on the right is a quantile-quantile plot of model residuals against theoretical quantiles of the Normal distribution.

correlate strongly with transcriptional activity [90, 53]. Here, we use experimental data from the genome-wide map of nucleosome acetylation and methylation published in [78] as an additional set of predictors underlying observed variation in yeast gene expression. The data consist of 19 extra variables including eight sets of histone modifications, measured under both normal conditions and following oxidative stress. This set also includes *in vivo* TFBS binding data for the TFs GCN4 and GCN5. We refer to these predictors as NUC.

The results were obtained using $\gamma_1 = (1, 1, 500)$, $\gamma_2 = (2, 1, 500)$, $\gamma_3 = (3, 1, 500)$ and $\gamma_4 = (2, 2, 500)$ and are presented in Table 3.5. Again, GEMULA identifies TFBS motifs that are known to be crucial for the transcriptional regulation of heat shock responsive genes. For instance, the GEMULA M2 model fitted using the MRM + NUC predictors contains the predictors MSN2, MSN4, ROX1, HSF1 and CST6. These 5 are among the 12 TFs that regulate gene regulatory modules of heat shock genes identified by Wu and Li in [117]. Furthermore, we again find that GEMULA models containing interactions between predictors outperform models with only main effects. Another noteworthy conclusion is that inclusion of the additional NUC predictors consistently leads to models with higher $\bar{R}^2_{cv}$s.

We further investigated the fit of the GEMULA M2 model for the yeast heat shock data using the TRAP 666 + NUC predictors, the model with the highest $\bar{R}^2_{cv}$, by evaluating two diagnostic plots. The plots are drawn in Figure 3.9. There is no clear pattern in the plot of the residuals against the fitted values in Figure 3.9 that would suggest strong heteroscedasticity or inadequacies of the linear model due to strong non-linear effects. This observation further supports our choice of a gene expression modeling strategy based on the linear model. Although from the qq-plot it is apparent that distribution of the residuals of the fitted model has somewhat heavier tails than the normal distribution, it does not deviate dramatically from normality. Overall, we conclude that the fit of the model inferred using GEMULA for the

observed expression at 20 minutes following heat shock is remarkably good.

Timepoint	Model	Predictors	Model type	Model P	Model T	$\bar{R}^2_{cv}$
49 mins	GEMULA	TRAP MRM	M1	30	30	0.13
49 mins	GEMULA	TRAP MRM	M2	34	81	0.28
49 mins	GEMULA	TRAP MRM	M3	15	35	0.14
49 mins	GEMULA	TRAP MRM	M4	17	54	0.15
49 mins	GEMULA	TRAP 666	M1	105	70	0.27
49 mins	GEMULA	TRAP 666	M2	57	110	0.36
49 mins	GEMULA	TRAP 666	M3	27	28	0.23
49 mins	GEMULA	TRAP 666	M4	30	57	0.25
49 mins	GEMULA	Dictionary	M1	199	199	0.40
49 mins	GEMULA	Dictionary	M2	34	84	0.34
49 mins	GEMULA	Dictionary	M3	15	63	0.18
56 mins	GEMULA	TRAP MRM	M1	39	39	0.20
56 mins	GEMULA	TRAP MRM	M2	34	67	0.34
56 mins	GEMULA	TRAP MRM	M3	15	30	0.27
56 mins	GEMULA	TRAP MRM	M4	17	26	0.26
56 mins	GEMULA	TRAP 666	M1	143	102	0.38
56 mins	GEMULA	TRAP 666	M2	54	45	0.36
56 mins	GEMULA	TRAP 666	M3	26	23	0.29
56 mins	GEMULA	TRAP 666	M4	29	44	0.36
56 mins	GEMULA	Dictionary	M1	170	170	0.48
56 mins	GEMULA	Dictionary	M2	34	77	0.41
56 mins	GEMULA	Dictionary	M3	15	82	0.26

Table 3.3: Comparison of GEMULA models inferred using different sets of predictors.

Timepoint	Model	Predictors	Model type	Model P	Model T	$\bar{R}^2_{cv}$
49 mins	MARS	TRAP MRM	$\kappa = 1, \lambda = 2$	26	31	0.07
49 mins	MARS	TRAP MRM	$\kappa = 1, \lambda = 3$	19	28	-0.03
49 mins	MARS	TRAP MRM	$\kappa = 1, \lambda = 4$	11	17	0.07
49 mins	MARS	TRAP MRM	$\kappa = 2, \lambda = 3$	30	57	-0.44
49 mins	MARS	TRAP MRM	$\kappa = 2, \lambda = 5$	20	23	-0.25
49 mins	MARS	TRAP MRM	$\kappa = 2, \lambda = 10$	11	9	-29622
49 mins	MARS	Dictionary	$\kappa = 1, \lambda = 2$	26	31	0.07
49 mins	MARS	Dictionary	$\kappa = 1, \lambda = 3$	20	22	0.03
49 mins	MARS	Dictionary	$\kappa = 1, \lambda = 4$	18	20	0.10
49 mins	MARS	Dictionary	$\kappa = 2, \lambda = 3$	26	32	-2.38
49 mins	MARS	Dictionary	$\kappa = 2, \lambda = 5$	25	31	-0.05
49 mins	MARS	Dictionary	$\kappa = 2, \lambda = 10$	13	10	-0.05
56 mins	MARS	TRAP MRM	$\kappa = 1, \lambda = 2$	17	26	0.04
56 mins	MARS	TRAP MRM	$\kappa = 1, \lambda = 3$	15	23	0.08
56 mins	MARS	TRAP MRM	$\kappa = 1, \lambda = 4$	12	15	0.10
56 mins	MARS	TRAP MRM	$\kappa = 2, \lambda = 3$	28	46	-618
56 mins	MARS	TRAP MRM	$\kappa = 2, \lambda = 5$	21	26	-0.34
56 mins	MARS	TRAP MRM	$\kappa = 2, \lambda = 10$	8	8	-0.02

Table 3.4: Comparison of MARS models inferred using different sets of predictors.

Predictors	Model	Model P	Model T	$\bar{R}^2_{cv}$
TRAP MRM	GEMULA M1	42	42	0.38
TRAP MRM	GEMULA M2	31	71	0.46
TRAP MRM	GEMULA M3	14	16	0.39
TRAP MRM	GEMULA M4	15	36	0.41
NUC	GEMULA M1	18	18	0.51
NUC	GEMULA M2	19	49	0.62
NUC	GEMULA M3	14	91	0.61
NUC	GEMULA M4	15	60	0.63
TRAP MRM + NUC	GEMULA M1	41	41	0.60
TRAP MRM + NUC	GEMULA M2	31	61	0.65
TRAP MRM + NUC	GEMULA M3	14	63	0.67
TRAP MRM + NUC	GEMULA M4	15	98	0.69
TRAP 666	GEMULA M1	75	75	0.55
TRAP 666	GEMULA M2	31	62	0.58
TRAP 666	GEMULA M3	14	46	0.54
TRAP 666	GEMULA M4	15	70	0.56
TRAP 666 + NUC	GEMULA M1	46	46	0.65
TRAP 666 + NUC	GEMULA M2	31	78	0.70
TRAP 666 + NUC	GEMULA M3	14	43	0.68
TRAP 666 + NUC	GEMULA M4	15	62	0.69

Table 3.5: Comparison of candidate models fitted using GEMULA on yeast heat shock gene expression data with different sets of predictors.

3.6 Application of GEMULA

To illustrate the potential of GEMULA as a method to analyze *mammalian* data, here we present the results of application of GEMULA to a gene expression time-course dataset of cultured F11 cells profiled at several time-points following Forskolin stimulation. Recall from Section 1.2 that F11 cells provide a good *in vitro* model for the transcriptional regulation of DRG regeneration *in vivo*. Upon stimulation with Forskolin, F11 cells acquire a neuronal phenotype which results in the outgrowth of neurites. Since it is a unicellular system, gene expression changes are more homogeneous and less complex than gene expression from *in vivo* samples of neuronal tissue where cells are in a complex and heterogeneous cellular environment. We therefore expect that the F11 gene expression dataset contains valuable information on transcriptional regulation underlying growth of neurites.

Microarray gene expression analysis

F11 cells were incubated in low-serum medium (DMEM with 0.5% FCS and antibiotics) for three hours and then stimulated with 10μM Forskolin for 0, 2, 4, 24 and 48 hours and total RNA was isolated using Trizol reagent. The RNA extracted from the cultured cells at each time-point was split into three batches to perform three technical replicates, i.e. hybridizations to three different microarray chips. RNA samples were amplified, labeled and hybridized to Agilent 4x44K Rat Whole-Genome expression arrays using standard Agilent protocols. The total number of microarray hybridizations (arrays) for the whole experiment is $4 \times 3 = 12$. Arrays were scanned using an Agilent scanner and data were read using Agilent Feature Extraction software. Array data were further processed using the R packages `bioconductor` (Gentleman *et al.* [36]) and `limma` (Linear Models for Microarray Data) (Smyth [95], Ritchie *et al.* [83]) for EdwardŠs background subtraction and loess normalization. We use the Bayesian Analysis of Time Series (BATS) method developed by Angelini *et al.* [6] to find probes on the array that are differentially expressed in response to Forskolin stimulation with respect to control (the expression at time-point 0). This Bayesian method designed for short replicated time-series was also used in the analysis of the rat DRG gene expression data in Section 2.2. For probes identified by BATS as significantly differentially expressed, we consider the log-fold ratio of expression in treatment versus control and average over the three technical replicates. We refer to the genes that correspond to these probes as *Forskolin responsive* genes. The vectors of gene expression of Forskolin responsive genes at 4 time-points after stimulation are our response variables of interest.

We consider two different sets of predictors. Both sets are derived from the same set of non-redundant vertebrate TFBS PSSMs from TRANSFAC [68] Professional, Release 11.1. Predictors in the first set, referred to as 'Counts TF11', represent counts of occurrences of predicted TFBSs in rat gene regulatory DNA sequences. The Counts TF11 predictors are derived from the rat TFBS annotation dataset described in Section 2.5. Here, we use the count of all occurrences of a TFBS in the entire regulatory DNA sequence of a gene as a predictor, instead of the binary variables indicating presence versus absence of the TFBS that are used by LLM3D (see Chapter 2). Predictors in the second set, referred to as TRAP TF11, represent binding affinities obtained with TRAP [85].

3.6.1 Results

We applied GEMULA to analyze F11 gene expression data and compare models fitted by GEMULA with models fitted using MARS. The results that we report were obtained using $\gamma_1 = (1, 1, 800)$, $\gamma_2 = (2, 1, 800)$ and $\gamma_3 = (3, 1, 800)$. Prior to running the MARS model selection algorithm, we selected the 50 predictors most strongly associated to the observed gene expression univariately, in the same way as in our analysis of the yeast data described in Section 3.5. The results obtained for the observed gene expression at 48 hours after stimulation, are presented in Table 3.6. The tables contain similar descriptive statistics on the fitted models to those we presented in Section 3.5.

Model	Predictors	Model type	Model P	Model T	$\bar{R}^2_{cv}$
GEMULA	TRAP TF11	M1	55	55	0.15
GEMULA	TRAP TF11	M2	38	60	0.17
GEMULA	TRAP TF11	M3	14	19	0.14
GEMULA	Counts TF11	M1	49	48	0.15
GEMULA	Counts TF11	M2	37	69	0.17
MARS	TRAP TF11	$\kappa = 1$, $\lambda = 2$	24	37	0.07
MARS	TRAP TF11	$\kappa = 1$, $\lambda = 4$	7	9	0.07
MARS	TRAP TF11	$\kappa = 2$, $\lambda = 3$	27	50	-0.19
MARS	TRAP TF11	$\kappa = 2$, $\lambda = 10$	11	10	0.07

Table 3.6: Comparison of models fitted using GEMULA and MARS for F11 gene expression data at 48 hours following Forskolin stimulation.

The differences between the results obtained using TRAP TF11 and Counts TF11 predictors are marginal. Both sets are able to "explain" about 17% of the observed expression, with a similar number of predictors and model terms. GEMULA models that include pairwise interactions between predictors fit better in terms of $\bar{R}^2_{cv}$ than models that only include main effects. Although the differences in $\bar{R}^2_{cv}$ between the M1 and M2 models fitted by GEMULA are small, the M2 models use less input predictors, but still fit the observed data better. The M3 model fitted using the TRAP TF11 predictors uses even less input predictors and model terms. Although the differences in number of predictors and model terms for models M1, M2 and M3 are considerable, the differences in terms of $\bar{R}^2_{cv}$ are small. We note that models fitted using MARS perform considerably worse in terms of $\bar{R}^2_{cv}$ than the models fitted with GEMULA. The default GCV penalty per knot for MARS when $\kappa = 2$ is $\lambda = 3$. Use of this value consistently resulted in models with negative $\bar{R}^2_{cv}$s on all time-points (data not shown), clearly indicating overfit. Increasing this penalty results in a model with positive $\bar{R}^2_{cv}$, but lower than the M2 model fitted using GEMULA.

In Table 3.7 we show a comparison of models fitted by GEMULA based on the TRAP TF11 predictors for *all* four time-points. We present results for the TRAP TF11 predictors only.

Time	Model type	Model P	Model T	$\bar{R}^2_{cv}$
2 hours	M1	43	42	0.07
2 hours	M2	40	154	0.13
2 hours	M3	16	48	0.10
4 hours	M1	34	34	0.03
4 hours	M2	39	86	0.05
4 hours	M3	16	78	0.05
24 hours	M1	33	33	0.13
24 hours	M2	38	52	0.13
24 hours	M3	16	33	0.12
48 hours	M1	55	55	0.15
48 hours	M2	38	60	0.17
48 hours	M3	14	19	0.14

Table 3.7: Comparison of models fitted by GEMULA for F11 gene expression data for all four time-points.

Results obtained with the Counts TF11 predictors are again very similar. From Table 3.7, it is interesting to note that the amount of expression variation that can be "explained" by TFBS motifs occurring in gene regulatory sequences seems to vary in time. For the 2 hour time-point, shortly after Forskolin stimulation, the $\bar{R}^2_{cv}$ for the GEMULA model that contains interactions is almost twice as large as the GEMULA model that only includes main effects. At four hours, the $\bar{R}^2_{cv}$s are considerably lower than $\bar{R}^2_{cv}$s obtained for the 2 hours and for the 24 hours and 48 hours time-points. When we examined the expression profiles of the regulated genes, we noticed clear differences between the expression at the first two time-points immediately following Forskolin stimulation and the two "late" time-points. This observation together with the results of Table 3.7 prompted us to further investigated the possible underlying time-dependent activity of transcription factors and the presence of an "early" and a "late" transcriptional response.

Separation and analysis of early and late transcriptional changes

Das *et al.*[23] noted that inclusion of a set of "background" genes, i.e. genes that are not actually regulated under the experimental condition of interest, adversely affects the overall fit of their models, because the measured gene expression for such genes constitutes only noise. The set of differentially regulated genes identified by BATS consists of genes that have a significantly altered expression in treatment with respect to control in *at least* one time-point, but not necessarily at *all* time-points. From a biological point of view it makes sense to distinguish between "early responsive" and "late responsive" genes, because interactions

between TFs and their target genes are known to be condition-specific and time-dependent. Therefore, we performed a principal component analysis on the matrix containing gene

Cluster	Sign of PC1 coef	Sign of PC2 coef	Number of genes
Early up	+1	+1	339
Late up	+1	−1	556
Late down	−1	+1	598
Early down	−1	−1	336

Table 3.8: Principal component based clustering of gene expression of Forskolin responsive genes in F11 cells.

expression of the Forskolin responsive genes to see if indeed the observed variation in expression can be further divided into biologically interesting patterns. For each gene, we use the signs of the coefficients corresponding to the first and second principal component to further define four homogeneous gene expression clusters. The first principal component seems to identify the main direction of the induced gene expression changes and can be used to distinguish between predominantly up- versus down-regulated genes, whereas the second principal component discriminates between genes that show an early and a late response. Based on these observations, we suggestively name the 4 resulting clusters "early up", "late up", "late down" and "early down", see Table 3.8. A plot of the expression averaged over all genes in each of the 4 clusters can be found in Figure 3.10.

We applied GEMULA again, but this time we distinguished between early responsive genes, consisting of all genes from both the "early up" and "early down" clusters, and late responsive genes, consisting of all genes from both the "late up" and "late down" clusters. At each time-point we fitted models with GEMULA using either set of genes. If the early responsive genes are transcriptionally regulated at the early, *but not* the late time-points, we expect the models fitted at the late time-points using data from the early responsive genes to have considerably lower $\tilde{R}^2_{cv}$s and vice versa. We indeed find that this is the case. Moreover, this time the fitted models have notably higher $\tilde{R}^2_{cv}$s than the models in Table 3.7. We present the results in Table 3.9. From these results we conclude that the experimental data suggest the presence of two separate waves of transcriptional changes.

Of particular interest biologically are the sets of TFs that are associated to early and late gene expression changes. Therefore, we consider some of the predictors that are present in the models selected by GEMULA in Table 3.10. For each predictor, this table includes a logo representation of the DNA sequences that are recognized by the corresponding TF. The GEMULA M3 model for the expression at 2 hours for the early responsive genes contains the predictors V.CREB.Q4.01, V.VJUN.01 and V.CEBPDELTA.Q6, among others. The V.CREB.Q4.01 motif represents the TFBS of CREB, a cAMP-inducible TF. Activation of CREB is known to be induced by Forskolin stimulation of F11 cells and Gao *et al.* [32] have shown that activated CREB is sufficient to promote spinal axon regeneration. The V.VJUN.01 motif, also known as AP-1, represents binding sites for dimers of transcription factors from the Jun,

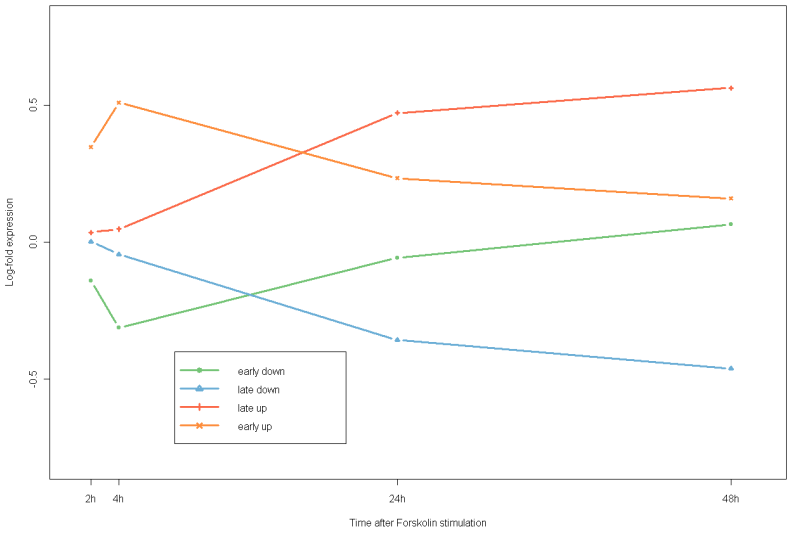

Figure 3.10: Plot of average gene expression of clusters of Forskolin responsive genes in F11 cells at several time-points following Forskolin stimulation.

Fos, and Atf families of DNA binding proteins. Several studies (Herdegen *et al.*[45, 46], Raivich *et al.* [81]) have implicated the TF c-Jun, which is known to bind to AP-1 sites, to successful axonal regeneration after injury. In contrast, the GEMULA M2 model for the expression at 48 hours for the late responsive genes contains the predictors V.CETS1P54.01, V.EBF.Q6, V.POU6F1.01, V.PPAR.DR1.Q2 and V.PPARA.01, among others. We already experimentally validated a role for PPARγ, which binds to the V.PPARA.01 motif, in the regulation of genes that are involved in neuron differentiation in Section 2.3. A study by Hippenmeyer *et al.* [48] reveals a role for two ETS TFs, which bind to the V.CETS1P54.01 motif, in neuronal differentiation of DRG neurons, whereas Bacon *et al.* [8] implicate putative ETS binding sites in the regulation of galanin and other axotomy-responsive genes in the DRG. The only overlap in predictors between the two GEMULA models considered in Table 3.10 is V.E2F.Q6.01, which represents binding sites of transcription factors from the E2F family. Members of this family are well-known regulators of the cell cycle. Altogether, we conclude that the obtained results are biologically plausible and suggest that dynamic and combinatorial activity of TFs underlie the observed early and late transcriptional changes in F11 cells in response to Forskolin stimulation. We will analyze the relative importance of the predictors in the selected models in the next chapter.

		Early responsive genes			Late responsive genes		
Time	Model type	Model P	Model T	$\bar{R}^2_{cv}$	Model P	Model T	$\bar{R}^2_{cv}$
2h	M1	30	30	0.14	8	8	-0.00
2h	M2	31	72	0.22	36	53	0.06
2h	M3	14	47	0.25	16	27	0.03
4h	M1	16	16	0.08	0	0	0
4h	M2	31	75	0.14	36	53	0.03
4h	M3	14	39	0.07	16	16	0.03
24h	M1	50	50	0.01	39	39	0.25
24h	M2	31	80	0.11	36	63	0.24
24h	M3	14	20	0.02	15	27	0.23
48h	M1	5	5	-0.01	44	44	0.25
48h	M2	31	85	0.11	35	52	0.27
48h	M3	14	60	0.04	16	37	0.24

Table 3.9: Comparison of models fitted using `GEMULA` for early and late Forskolin responsive genes in F11 cells at all four time-points. Columns 3-5 correspond to models fitted for the early responsive genes and columns 6-8 to models for the late responsive genes.

Time-point	Model	TFBS ID	Motif logo
2 hours	M3	V.VJUN.01	
2 hours	M3	V.CREB.Q4.01	
2 hours	M3	V.CEBPDELTA.Q6	
48 hours	M2	V.CETS1P54.01	
48 hours	M2	V.EBF.Q6	
48 hours	M2	V.PPAR.DR1.Q2	
48 hours	M2	V.PPARA.01	
48 hours	M2	V.POU6F1.01	

Table 3.10: TFBS motif logos of predictors in models selected by `GEMULA`.

3.7 Discussion

Regression models are valuable tools for inference of transcriptional gene regulatory interactions from using gene expression and DNA sequence data. In particular they allow identification of interactions between predictors that underlie observed patterns of variation in gene expression under different experimental conditions or across time. The success of a regression based approach depends on appropriate choices for the type of model and the predictors used as input. This was first demonstrated by Das *et al.* [23] who proposed a strategy that uses MARS as core regression method [23, 24, 25]. In [23] Das *et al.* claim that their MARSMOTIF algorithm, which allows modeling of synergistic interactions between predictors, is approximately 1.5 to 3.5 times more accurate than the REDUCE method proposed by Bussemaker *et al.* [17], which is based on a linear model. The comparison of MARSMOTIF and REDUCE is based on an R^2-like $\Delta\chi^2$ statistic and no cross-validation was performed. A comprehensive comparison is lacking. The results we present in this chapter show that similar synergistic interactions as considered by Das *et al.* in [23, 24] can be modeled using linear models. We find that GEMULA, a linear model based approach combined with a powerful procedure to generate and select among a wide range of models with varying degree of interactions, leads to biologically plausible models with good fit. Furthermore, cross-validation indicates that GEMULA models suffer substantially less from lack-of-fit than MARS models fitted on the same data.

In order to build models that are biologically useful and interpretable, the availability of relevant biological predictors used as input are crucial. The TRAP predictors we consider in this chapter are non-degenerate and represent *in silico* predicted binding affinities of TFs. Using yeast data, we showed that GEMULA in combination with TRAP predictors successfully identifies interactions between known cell cycle regulating TFs such as MBP1, MCM1, SWI5, SWI6 and FKH2 that underlie the observed periodic expression patterns of cell cycle genes in yeast. The TRAP predictors are real-valued and have an interpretation that is closer to experimentally measured TF-DNA binding profiles as obtained for instance with ChIP-chip assays, than non-degenerate motif representations that use exact words. Models that use TRAP predictors have a clear interpretation which facilitates the step toward biological validation. A strategy with great potential is the use of other biologically important predictors of gene expression in addition to TRAP predictors. Examples of such predictors include experimental TF binding data and data on chromatine modifications and nucleosome occupancy which are important determinants of promoter accessibility. Our analysis of yeast heat shock gene expression data shows that GEMULA can integrate different sources of experimental data resulting in models with higher "explanatory value" than models that consider only predictors that represent TFBSs.

Even in the absence of additional predictors, GEMULA can be used to analyze *mammalian* gene expression data. We applied GEMULA to identify TFs associated to observed patterns of early and late gene expression changes in F11 cells in response to Forskolin stimulation. The observed fit in terms of $\bar{R}^2_{cv}$ of the resulting models in Section 3.6 is considerably less than for GEMULA models inferred from yeast heat shock data in Section 3.5.2 where additional predictors are available. To put this into perspective, we compare our results to the work by Das *et al.* [24] who applied MARS to model gene expression in several human tissues. Their

work on regression based modeling of mammalian gene expression data is the most closely related to our work we are aware of. They report an average $\Delta\chi^2$ of 21.7% for models fitted on a time-course dataset of human cell cycle gene expression data and an average $\Delta\chi^2$ of 24.4% for tissue-specific gene expression data. In comparison, the GEMULA M3 model for the F11 gene expression at 2 hours for the early responsive genes has a $\Delta\chi^2$ of 32% and the GEMULA M2 model for expression of late responsive genes at 48 hours has a $\Delta\chi^2$ of 30%. In the near future, as more genome-wide experimental data will become available, we believe that GEMULA will prove to be a useful tool in bridging the gap in understanding of transcriptional networks between yeast and the more complex mammalian systems.

FOUR

ESTIMATION OF VARIABLE IMPORTANCE

The regression models we studied in the previous chapter are used to model observed variation in gene expression as a function of many predictor variables simultaneously. When modeling real experimental data, it is generally not possible to select a single model and a corresponding set of predictors that clearly achieves the best fit to the observed data. Furthermore, selected models do not always contain parameters that can be used to quantify the importance of the individual predictors in the model. In this chapter, we use the statistical framework of estimation of variable importance to define variable importance as a parameter of interest and study two different estimators of this parameter in the context of modeling gene expression data. On yeast data we show that the resulting parameter has a biologically appealing interpretation. We apply the variable importance methodology on mammalian gene expression data to gain insight into the temporal activity of TFs that underly gene expression changes in F11 cells in response to Forskolin stimulation.

4.1 Introduction

The regression models we have considered in the previous chapter are used to model gene expression as a function of many predictors simultaneously. When only a modest amount of, typically noisy, data is available, it is not realistic to expect that one can find a single best model among candidate models that clearly represents the "best" model. Still, even when the observed global fit of the models is not completely satisfactory, the inferred models provide us with valuable biological insights. In particular, we demonstrated in Chapter 3 that models fitted by GEMULA do allow us to *identify* predictors associated to variation in gene expression. In practice, however, there are usually several candidate models that fit almost equally well, as we also saw in our data analysis. These models may contain different, partially overlapping, sets of predictors. Moreover, the predictors typically occur in many model terms and a single term that can be interpreted as a marginal effect, such as a main effect term, is often lacking. In this chapter, our goal is to estimate the *marginal importance* of each predictor individually. The approach we present here is especially suited for situations in which ordinary least squares regression does not provide suitable models. It assumes that an appropriate method to fit a model for a given gene expression response of interest based on a set of relevant predictors is given and as such extends the work of the previous chapter.

From a practical point of view, quantifying the marginal importance of predictor variables is an important next step in the interpretation of the results of the previous chapter. In order to do so, we define this importance as a parameter of interest and consider estimators of this parameter. This means we shift the focus from estimating a model for Y based on many predictors to estimating the importance of a single variable in such a model. It is important to bear in mind that the model selection procedure implemented in GEMULA is a procedure that trades off bias and variance to fit a good model for Y based on (a subset of) $X_1, \ldots, X_p$. When the true interest is in the marginal importance of a *single* variable, inference regarding this parameter based directly on the inferred model may be more biased than necessary (Van der Laan and Rubin [109]). In this chapter we use the framework of statistical inference for variable importance developed by Van der Laan [110] and show how it can be applied to define and estimate variable importance within the context of the models we developed in Chapter 3.

The remainder of this chapter is organized as follows. In Section 4.2 we give a definition of a variable importance measure (VIM) that makes sense within the context of the models we considered in Chapter 3 and introduce two different estimators. In Section 4.3 we study the behavior of these estimators in a simulation study. We show that the VIM we define represents a parameter that has an interesting biological interpretation by analyzing yeast gene expression in Section 4.4. Finally, we apply the VIM methodology to study the involvement of transcriptional regulators in determining gene expression of neuronal outgrowth-associated genes in Section 4.5. We conclude with a discussion in Section 4.6.

4.2 Methods

4.2.1 Marginal variable importance as a real-valued parameter

Suppose we observe a set of p predictors $X_1,\ldots,X_p$ and a response variable Y, all vectors of length n. We are interested in the marginal variable importance of X_j in determining Y, in a model where also possibly confounding predictors $X^*_{-j} = \{X_1,\ldots,X_{j-1},X_{j+1},\ldots,X_p\}$ may be related to Y. Hence, when we model the effect of variable j, for $j = 1\ldots,p$ we consider the other variables X^*_{-j} as nuisance variables. For notational convenience we fix j and let $Z = X_j$ and $X^*_{-j} = X^*$. Within the VIM framework proposed by Van der Laan [110], variable importance is modeled using a semi-parametric model that describes the effect of Z,X^* on Y as

$$\mathbb{E}(Y|Z,X^*) = m(Z,X^*|\beta) + g(X^*), \tag{4.1}$$

where $g(X^*)$ is an unspecified function of X^* and m is an *a priori* given model, which models the effect

$$m(Z = z,X^*|\beta) = \mathbb{E}[Y|Z = z,X^*] - \mathbb{E}[Y|Z = 0,X^*], \tag{4.2}$$

for all z. Based upon this specification, the following general definition of *marginal variable importance* ψ is suggested.

Definition Let models $\mathbb{E}(Y|Z,X^*)$ and $m(Z,X^*|\beta)$ as specified in (4.1) and (4.2) be given. The *marginal variable importance* (VIM) of variable Z at $Z = z$, denoted by $\psi(z)$, is defined as

$$\psi(z) = \mathbb{E}_{X^*}[m(z,X^*|\beta)]. \tag{4.3}$$

In this chapter, we assume a linear model $m(Z,X^*|\beta) = \beta_j Z$ to model linear marginal effects. Furthermore, we consider $\psi = \psi(1)$ as the parameter of interest. The interpretation of this parameter is the expected change in Y for a unit change in Z while holding all other predictors fixed at their original values.

Example Let us consider an example. Suppose we have the following multiple linear regression model relating a response variable Y to a set Z,X^* of predictors

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_j Z + \ldots + \beta_p X_p + \epsilon.$$

Within the framework introduced above, we write this as $\mathbb{E}(Y|Z,X^*) = m(Z,X^*|\beta) + g(X^*)$, where $m(Z,X^*|\beta) = \beta_j Z$ and

$$g(X^*) = \beta_0 + \beta_1 X_1 + \ldots + \beta_{j-1} X_{j-1} + \beta_{j+1} X_{j+1} + \ldots + \beta_p X_p.$$

In this case, the variable importance parameter is given by

$$\psi(z) = \mathbb{E}[Y|Z = z,X^*] - \mathbb{E}[Y|Z = 0,X^*] = \beta_j z,$$

and we focus on inference of $\psi(1) = \beta_j$.
***

For this example the function $g(X^*)$ consists exclusively of additive main effects of the variables X^*. In general, more complex functions $g(X^*)$ can be considered. Our practical interest is in estimating the importance of biological predictors associated to gene expression and the general VIM definition introduced here allows us to use our GEMULA algorithm developed in the previous chapter to estimate $g(X^*)$ when analyzing real gene expression data.

4.2.2 Estimation of variable importance

Here, we discuss two methods to estimate the VIM parameter as defined in Equation (4.3). As is clear from the model specification in Equation (4.1), estimation of VIM requires estimation of $\mathbb{E}(Y|Z,X^*)$. For a given predictor Z, we adapt GEMULA to fit a model of the form

$$\mathbb{E}(Y|Z,X^*) = \beta_j Z + g(X^*), \tag{4.4}$$

where the allowed candidate terms $g(X^*)$ are determined by GEMULA through the specification of $\gamma = (\gamma_1, \gamma_2, \gamma_3)$. The model selected by GEMULA is used to produce a *penalized* VIM (pVIM) estimate of the parameter β_j in Equation (4.4). We denote this pVIM estimate by $\hat{\psi}_p$.

Targeted variable importance (tVIM)

Since the pVIM estimate is based on a L_1-penalized estimate of $\mathbb{E}(Y|Z,X^*)$ where the optimal shrinkage parameter was selected to obtain a good model for Y based on all predictors, the pVIM estimate may be more biased than necessary (Van der Laan and Rubin [109]). Van der Laan and Rubin [109] propose targeted maximum likelihood estimation to obtain a targeted VIM (tVIM) estimate. The tVIM estimate is obtained by updating the initial regression estimate $\mathbb{E}(Y|Z,X^*)$ in a direction which targets the parameter ψ of interest. The update takes into account the effect of confounders X^* on Z and therefore requires estimation of $G(X^*) = \mathbb{E}(Z|X^*)$. It is shown in [109] that when either $\mathbb{E}(Y|Z,X^*)$ or $\mathbb{E}(Z|X^*)$ are specified correctly, the tVIM estimator is consistent and asymptotically normal. Furthermore, if both models are specified correctly, the tVIM estimate is efficient. In practice, the overall quality of the estimates depends on good estimates of $\mathbb{E}(Y|Z,X^*)$ and $\mathbb{E}(Z|X^*)$. We point out that the outlined VIM framework is general and that $\mathbb{E}(Y|Z,X^*)$ and $\mathbb{E}(Z|X^*)$ may be estimated using any statistical method deemed appropriate in the given context of the application. We use the GEMULA algorithm with context dependent parameters $\gamma = (\gamma_1, \gamma_2, \gamma_3)$ to obtain VIM estimates when estimating variable importance of predictors associated to variation in gene expression. Let an initial fit of a model M_0 for $\mathbb{E}(Y|Z,X^*)$ and a fit of a model M_G for $\mathbb{E}(Z|X^*)$ be given. Below, we give the steps required for the computation of tVIM. For more details, we refer to [110, 109].

1. Calculate a covariate $r(Z,X^*) = Z - \hat{Z}^{M_G}$ from the fitted model M_G for $\mathbb{E}(Z|X^*)$.

2. Compute the vector of fitted response values $\hat{Y}^{M_0}$ according to the fitted model M_0 for $\mathbb{E}(Y|Z,X^*)$.

3. Regress Y on $r(Z, X^*)$ using $\hat{Y}^{M_0}$ as an *offset* and denote the estimated regression coefficient by $\hat{\epsilon}$. An offset is a term that can be added to a linear model and that is treated as an *a priori* known term, for which no coefficient needs to be estimated. The offset is subtracted from the response prior to fitting. The estimate $\hat{\epsilon}$ can be obtained by standard OLS regression, using a model without an intercept term but *with* the mentioned offset.

4. Update the initial pVIM estimate to obtain the tVIM estimate $\hat{\psi}_t$ as

$$\hat{\psi}_t = \hat{\psi}_p + \hat{\epsilon}.$$

4.3 Simulation study

Because the VIM defined in (4.3) in Section 4.2.1 defines a linear effect, we use the pilot model that we introduced in Section 3.3.1 to compare the estimation of VIMs using the pVIM and tVIM estimators described in Section 4.2.1. The main purpose is to show the effect of the targeting step on the performance of the tVIM estimator and to get some insight into its behavior. Recall that the pilot model introduced in Section 3.3.1 was designed primarily to study models relating binding affinities of DNA binding TFs to observed variation in gene expression in a practically relevant context. Hence, simulations using this model provides us with perspective on the potential of VIM estimation for the identification and ranking (based on importance) of predictors associated to variation in gene expression. Here, we consider a data generating model that contains the linear main effects for all 33 predictors in the pilot model. We use the set of 123 `TRAP MRM` predictors as candidate predictors. Hence, only the 33 predictors present in the pilot model correspond to "truly important" predictors, i.e. predictors with a non-vanishing regression coefficient. For $j = 1, \ldots, 33$, we let β_j represent the parameter of interest, i.e. the true VIM parameter ψ_j corresponding to predictor j. The response variable Y is generated as

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_{33} X_{33} + \epsilon, \tag{4.5}$$

where $\epsilon = (\epsilon_1, \ldots, \epsilon_n) \sim \mathcal{N}(\mathbf{0}, \sigma^2 I_n)$. In the simulations we consider here, we fix the sample size at $n = 790$, motivated by our analysis of the 790 cell cycle genes in Section 3.3.1 and the marginal effect of varying the sample size in a realistic range we have observed in the simulation study there. We set $\sigma^2 = 0.26$, which corresponds to a setting with high noise variance. On data simulated according to model (4.5), we compare the performance of the pVIM and tVIM estimators, based on 200 independent simulation runs. We also compare the pVIM and tVIM estimates to estimates of the regression coefficients obtained with OLS. Hence, in each independent simulation run we record the estimated VIM according to the following three estimators.

1. **mOLS**. The first method we use to estimate the VIM for each candidate predictor j, for $j = 1, \ldots, 123$, is based on an OLS fit of the model

   $$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_{123} X_{123} + \epsilon. \tag{4.6}$$

   The estimate of β_j obtained from the resulting fit is recorded as the mOLS estimate of ψ_j.

2. **pVIM**. We use `GEMULA` with $\gamma = (1, 1, 123)$ to obtain the pVIM estimate. With this setting, the model selected by `GEMULA` corresponds to an L_1-penalized lasso fit of model (4.6). The estimate of β_j that is obtained as the estimated regression coefficient corresponding to predictor j in the model selected by `GEMULA` is recorded as the pVIM estimate of ψ_j.

3. **tVIM**. The tVIM estimate is obtained by applying the steps described in Section 4.2.2 to the pVIM estimate of ψ_j in the previous step. To estimate $\mathbb{E}(Z|X^*)$ we use `GEMULA`

with $\gamma = (1, 1, 15)$.

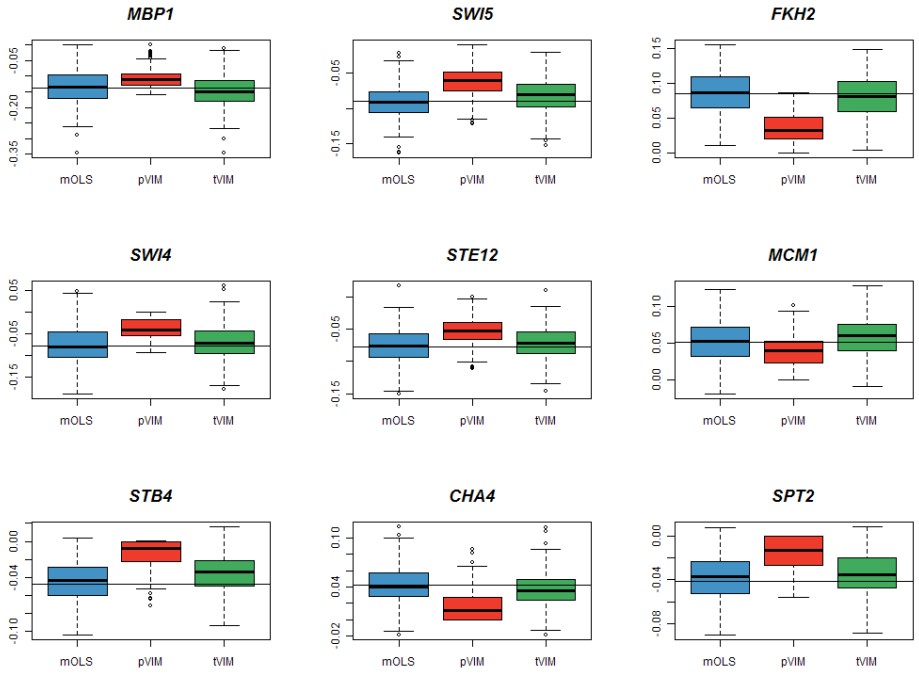

Figure 4.1: Boxplots of VIMs estimated using three different estimators for the 9 most important predictors in the simulation study.

The predictors can be ranked according to their true importance $|\beta_j|$. Figure 4.1 contains box plots of the estimated VIMs according to the three methods for the 9 highest ranking predictors. The horizontal line in each plot represents the true VIM ψ_j. The plots in Figure 4.1 clearly illustrate how the targeting step works. It moves the shrunken, low variance (but biased) pVIM estimate in the direction of the true value of the parameter. As such, the resulting targeted VIM estimate represents a compromise between pVIM and mOLS. On average, it has a lower bias than pVIM but a higher variance. The quality of an estimator is a function of both bias and variance and a common way to quantify the distance between an estimator and a parameter of interest being estimated is to compute the mean square error (MSE), or its square root (RMSE). Table 4.1 contains RMSEs calculated for each of the three different estimators based on 200 simulations. From this Table, we conclude that for these "most important" predictors, tVIM gives the most accurate estimates in terms of RMSE. Whereas we have seen in Section 3.3.3 that the models selected using lasso-AIC shrinkage from which the pVIM estimates are obtained are good if our interest is in explaining the variation estimating in the response variable Y, the estimated regression coefficients of the predictors in the models may be more biased than necessary. However, we note that this

Predictor	mOLS	pVIM	tVIM
MBP1	0.0585	0.0447	**0.0543**
SWI5	0.0251	0.0350	**0.0245**
FKH2	0.0315	0.0535	**0.0303**
SWI4	0.0436	0.0472	**0.0424**
STE12	0.0295	0.0332	**0.0283**
MCM1	0.0276	**0.0249**	0.0269
STB4	**0.0216**	0.0374	0.0233
CHA4	0.0223	0.0314	**0.0212**
SPT2	0.0215	0.0302	**0.0208**

Table 4.1: RMSEs of three different VIM estimators for the 9 most important predictors in the simulation study. The smallest RMSE is indicated in boldface.

is not necessarily so for *all* predictors. As the effects ψ_j become smaller, at some point the negative impact of the additional variance in the estimates introduced by the targeting step overcomes the benefits of the reduction in bias. For smaller effects, shrinking them toward zero results in lower RMSEs. Hence, for predictors with very small effects, the pVIM estimates are most accurate. This is illustrated in Figure 4.2 and Table 4.2.

Predictor	mOLS	pVIM	tVIM
MOT3	0.0221	0.0218	**0.0208**
GAL4	0.0251	**0.0220**	0.0229
PHO4	0.0361	**0.0226**	0.0324
RPH1	0.0216	0.0210	**0.0205**
MET28	0.0263	**0.0204**	0.0243
ASH1	0.0251	**0.0197**	0.0221
CAD1	0.0340	**0.0178**	0.0329
ARR1	0.0209	**0.0150**	0.0195
ACE2	0.0272	**0.0142**	0.0258

Table 4.2: RMSEs of three different VIM estimators for the 9 least important predictors in the simulation study. The smallest RMSE is indicated in boldface.

Hence, overall we conclude that both pVIM and tVIM provide good estimates of VIMs of interest on which rankings of marginal importance of predictors can be based. The main purpose of the simulations we present here is to illustrate the variable importance framework

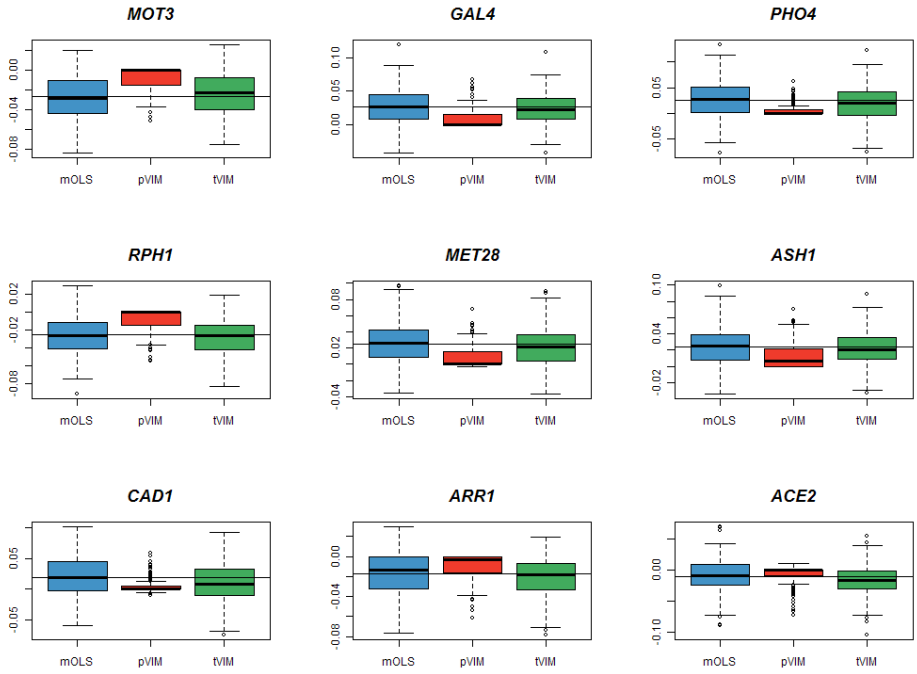

Figure 4.2: Boxplots of VIMs estimated using three different estimators for the 9 least important predictors in the simulation study.

within a clearly interpretable context and to characterize the effect of the *targeting* step on the pVIM estimates. Although in this simulation example mOLS appears to yield reasonable VIM estimates too, it almost never outperforms both pVIM and tVIM. Moreover, within the general variable importance framework outlined in Section 4.2.1, it is not a natural estimator to consider. This framework enables us to model variable importance in the context of modeling real experimental gene expression data with `GEMULA` or MARS we considered in Chapter 2.5. We consider pVIM and tVIM to be complementary and use them both to analyze real expression data. The relative usefulness of the VIM estimates and rankings of predictors obtained using pVIM and tVIM will become clear upon further validation and interpretation of the inferred results obtained on real gene expression data.

4.4 Validation on yeast gene expression data

In order to confirm that estimation of variable importance using pVIM and tVIM estimators yields biologically relevant parameters when applied to real experimental data, we apply the outlined variable importance approach to the yeast cell cycle gene expression data introduced in Section 3.3.1. This time, we give a comprehensive analysis of the relative importance of different TFs that are associated to the observed variation in gene expression and focus on the dynamic activity of the TFs in time. In their analysis of the 800 periodically expressed genes they identified as cell cycle regulated, Spellman *et al.* partitioned this set into five subsets based on the moment of peak expression during the cycle. In the following we use data from the entire set of experiments were α-factor arrest was used to synchronize the yeast cells. Expression was measured at 7 minute intervals up to 119 minutes after synchronization. Hence, the dataset we analyze consists of time-course gene expression profiles for all known yeast genes at 18 different time-points spanning two complete cell cycles. Figure 4.3 shows the average expression profiles of the 800 periodically expressed genes clustered by time of peak expression. In this plot the distinct cell cycle phases are indicated in boldface font. This plot clearly shows the periodicity of the gene expression response and the different

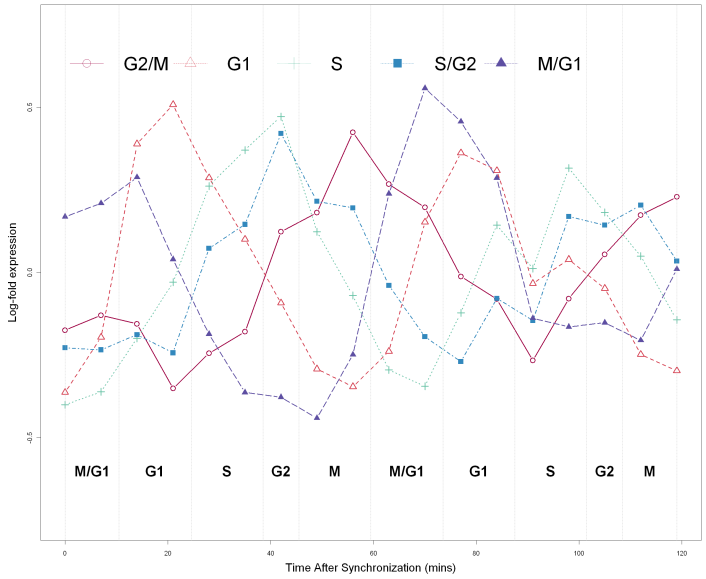

Figure 4.3: Observed gene expression across two complete cell cycles of 800 cell cycle regulated yeast genes, clustered by time of peak expression.

moments of peak expression of the different clusters of genes. Transcriptional regulation of cell cycle periodic genes has been studied intensively and analysis of different sources of experimental data has identified various TFs that underlie the periodic patterns of gene expression [107, 20, 118]. Cokus *et al.*[20] describe interactions between the primary or

canonical cell cycle regulators SWI4, SWI6, MBP1, FKH2, NDD1, MCM1, SWI5 and ACE2 which are known to form complexes and regulate phase transitions in the cycle in a serial fashion. Tsai *et al.* [107] identify a set of thirty putative cell cycle TFs. For nineteen of these there is strong evidence in the literature. The list of cell cycle TFs reported in [107] includes the eight canonical TFs discussed in [20]. In the canonical model of transcriptional regulation of the cell cycle, the different primary regulators activate their targets at the different phases (M/G1, G1/S and G2/M) in the cell cycle. We investigate whether we can reconstruct the activities of these canonical TFs by estimating their marginal variable importance for the different cell cycle phases.

In order to identify the TFs that control the expression of these genes, we rank candidate predictors based on their estimated marginal variable importance using the pVIM and tVIM estimators. We again use the TRAP MRM predictors (see Section 3.5) that are constructed using PFMs from 123 different yeast TFs derived from experimental binding data published by MacIsaac *et al.* [66]. We estimate $\mathbb{E}(Y|Z,X^*)$ using GEMULA with parameter $\gamma = (2,1,250)$. The resulting fitted model is used to produce the pVIM estimate of ψ_j. To compute the tVIM estimate of ψ_j, we estimate $G(X^*) = \mathbb{E}(Z|X^*)$ using GEMULA with $\gamma = (1,1,15)$ and update the pVIM according to the steps described in Section 4.2.2. Figure 4.3 shows a large cluster

Predictor	tVIM	tVIM rank	pVIM	pVIM rank
MBP1	0.203	1	0.199	1
STB1	0.092	2	0.081	2
SFP1	-0.082	3	-0.023	11
FKH2	-0.073	4	-0.061	3
HAC1	0.068	5	0.033	5
REB1	-0.056	6	-0.025	9
SKO1	0.055	7	0.031	6
ACE2	-0.051	8	-0.026	8
ASH1	-0.046	9	-0.028	7
AZF1	-0.044	10	-0.034	4
YAP3	-0.043	14	-0.025	10
DIG1	0.044	11	0.023	12

Table 4.3: Top ranked predictors by pVIM and tVIM. The response variable is observed gene expression of yeast cell cycle regulated genes 21 minutes after synchronization. The canonical cell cycle regulators are indicated in boldface and TFs belonging to the set of 19 known cell cycle TFs in [107] in italics.

of genes that peak 21 minutes following alpha synchronization, a time-point that lies within the G1 phase of the cell cycle. Table 4.3 lists the highest ranked predictors, ranked based on both pVIM and tVIM for this time-point. The top ranked predictors MBP1 and STB1 are both

known transcriptional activators of cell cycle genes during the G1 phase of the cycle in [107]. The positive value for the estimate of the VIMs of MBP1 and STB1 at this time-point indeed agree with their known role as activators of genes during G1. Table 4.3 also identifies the canonical regulators FKH2 and ACE2. Furthermore, the factor SFP1 is a known regulator of G2/M cell cycle transitions (note the negative sign of the estimated variable importance during G1) [18] and also ASH1 and DIG1 are implicated in regulation of cell cycle genes according to [107].

Predictor	tVIM	tVIM rank	pVIM	pVIM rank
MBP1	-0.125	1	-0.110	1
FKH2	0.089	2	0.040	6
SWI4	-0.079	3	-0.062	2
MCM1	0.076	4	0.056	4
STE12	-0.071	5	-0.061	3
SWI5	-0.068	6	-0.048	5
PHO4	0.045	7	0.002	18
ACE2	-0.044	8	0	NA
FKH1	-0.038	9	0	NA
RDS1	0.035	10	0.018	8
PDR3	0.024	22	0.018	7
PHD1	0.035	11	0.015	9
GCN4	0.001	50	0.014	10

Table 4.4: Top ranked predictors by pVIM and tVIM. The response variable is observed gene expression of yeast cell cycle regulated genes at 56 mins after synchronization. The canonical cell cycle regulators are indicated in boldface and TFs belonging to the set of 19 known cell cycle TFs in [107] in italics.

Another important gene expression pattern is due to genes that peak at the transition from G2 to M phase, corresponding roughly to the time-point 56 minutes after synchronization (see Figure 4.3). The top ranked predictors for this time point are listed in Table 4.4. We find high positive marginal importances of the canonical factors FKH2 and MCM1, both linked to the activation of M and G2/M cell cycle genes respectively according to [107]. Apart from MBP1 and MCM1, the top ranked predictors in Table 4.4 also include the canonical regulators SWI4, SWI5, ACE2 and FKH2. Note that FKH1, a TF that is part of the set of nineteen TFs with literature support for being important in cell cycle regulation according to Tsai *et al.* [107], and ACE2 can only be identified using tVIM. Also note that the crucial M phase regulator FKH2 ranks second in the tVIM list and only sixth in the list produced using pVIM. In contrast, we found no evidence in the literature for any specific cell cycle regulatory role for the TFs PDR3 and GCN4, which only receive high ranks according to pVIM. Together, these findings

illustrate the additional benefit of the targeting step and the tVIM estimator. The usefulness

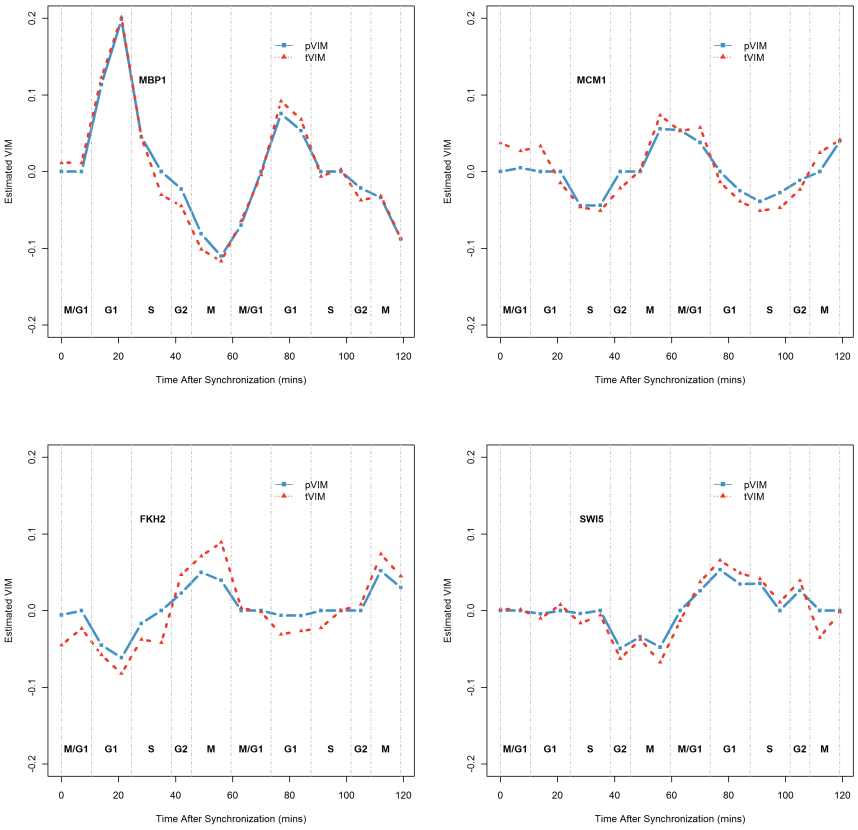

Figure 4.4: Plot of estimated VIMs of four canonical yeast cell cycle regulators across differents phases of the cell cycle.

of the variable importance parameter as defined by (4.1) and (4.3) is further demonstrated in Figure 4.4. This plot shows the estimated marginal variable importance of the canonical cell cycle TFs `MBP1`, `MCM1`, `FKH2` and `SWI5` as a function of time in the succesive stages of the cell cycle. Most prominent is the clearly periodically varying importance of `MBP1`, peaking in the `G1` phase. This is in good agreement with `MBP1`'s known role as activator of cell cycle genes at the transition from `G1` to `S` phase. Furthermore, the plots in Figure 4.4 suggest `MCM1` and `FKH2` as `G2/M` regulators and an involvement of `SWI5` in the `M/G1` transition. All of these findings are in agreement with what is known in the literature about the transcriptional effects of these TFs.

4.5 Estimation of VIM: an application

Here we apply the VIM methodology to estimate the variable importance of transcriptional regulators associated to early and late gene expression changes in F11 cells in response to stimulation with Forskolin. The data were introduced in Section 3.6. There, we showed that the entire set of Forskolin responsive genes can be further divided into groups of "early" and "late" responsive genes and we applied `GEMULA` to infer two models for the early responsive genes at the 2 hour and 4 hour time-point and two models for the late responsive genes at the 24 hour and 48 hour time-point. We again distinguish between these two groups. For the early responsive genes, we use `GEMULA` with $\gamma = (2, 1, 500)$ for the estimation of $\mathbb{E}(Y|Z, X^*)$ and for the late responsive genes, which is a bigger set, we use `GEMULA` with $\gamma = (2, 1, 700)$. For the estimation of $\mathbb{E}(Z|X^*)$ we use $\gamma = (2, 1, 110)$. The results for the

Predictor	tVIM	tVIM rank	pVIM	pVIM rank
V.CEBPDELTA.Q6	0.063	1	0.047	1
V.OCT1.03	0.056	2	0.031	3
V.PAX4.02	0.053	3	0.017	11
V.CIZ.01	0.051	4	0.02	8
V.YY1.Q6.02	-0.051	5	-0.012	14
V.CP2.02	0.047	6	0.026	5
V.CREB.Q4.01	0.043	7	0.039	2
V.AP1.Q4.01	0.041	8	0.02	9
V.DR1.Q3	-0.039	9	0	NA
V.LEF1.Q2.01	0.037	10	0.006	23
V.PBX.Q3	-0.035	11	-0.02	7
V.AREB6.02	-0.034	13	-0.027	4
V.VJUN.01	0.026	20	0.017	10
V.E2F.Q6.01	0.009	49	0.021	6

Table 4.5: Top ranked predictors by pVIM and tVIM. Response variable Y represents log-fold gene expression in cultured F11 cells of early responsive genes at 2h after Forskolin stimulation with respect to control.

first time-point at two hours following Forskolin stimulation are presented in Table 4.5. Note the high ranking of the binding site motifs `V.CREB.Q4.01` and `V.AP1.Q4.01`. We already discussed the role of the `V.CREB.Q4.01`, `V.AP1.Q4.01` and `V.VJUN.01` binding site motifs in driving gene expression in biological models in neuronal regeneration in Section 3.6. We report the results for two later time points in Table 4.6 and 4.7. According to these tables, there is a strong repression of genes by the known cell cycle regulator `E2F`. Since cell cycle arrest and neurogenesis are highly coordinated and interactive processes

Predictor	tVIM	tVIM rank	pVIM	pVIM rank
V.E2F.Q6.01	-0.123	1	-0.1	1
V.MYB.Q3	-0.081	2	-0.013	16
V.LRF.Q2	0.077	3	0	NA
V.AP1.Q4.01	0.066	4	0.016	9
V.COUP.DR1.Q6	0.057	5	0.015	10
V.GEN.INI3.B	0.056	6	0.008	20
V.E2A.Q2	0.056	7	0.018	6
V.EBF.Q6	0.054	8	0.033	3
V.OCT1.Q5.01	-0.05	9	-0.001	29
V.NKX3A.01	-0.045	10	-0.007	23
V.POU6F1.01	-0.045	11	-0.025	4
V.PPAR.DR1.Q2	0.043	12	0.017	7
V.PAX4.03	0.043	13	0.045	2
V.PPARA.01	0.042	14	0.015	11
V.P300.01	0.039	18	0.017	8
V.VDR.Q3	0.039	19	0.025	5

Table 4.6: Top ranked predictors by pVIM and tVIM. Response variable Y represents log-fold gene expression in cultured F11 cells of late responsive genes at 24h after Forskolin stimulation with respect to control.

(Ohnuma *et al.* [75]), the involvement of E2F in regulation of genes in Forskolin stimulated F11 cells is plausible. Among the top 10 ranked TFBS motifs at the 24 hours and 48 hours time-point are V.PPARA.01 and V.PPAR.DR1.Q2. The consensus sequence of the TFBSs corresponding to this motif is recognized by TFs from the family of *peroxisome proliferator-activated receptors* (PPARs). In Chapter 2, we predicted PPARs to regulate genes involved in neuronal differentiation based on analysis of *in vivo* gene expression data from rat DRG neurons in response to injury using LLM3D. We described the validation of the effect of PPARγ on regulation of genes involved in neuronal differentiation in Section 2.3. Our findings here provide further support for our claim that PPARγ is an important transcriptional regulator in neuronal regeneration. In addition to V.PPARA.01 in Table 4.6 and Table 4.7 we find V.EBF.Q6. This motif is bound by *early B-cell factor* (EBF) TFs. Garel *et al.* [34] find that EBFs are potentially involved in neuronal differentiation in the developing CNS. In [33], Dominguez *et al.* find that EBFs appear to be master controllers of neuronal differentiation and migration, coupling them to cell cycle exit and earlier steps of neurogenesis. A review by Liberg *et al.* [62] discusses the role of EBFs as regulators of differentiation in embryonic neural development. This review also describes interactions between *CCAAT/ enhancer-*

binding proteins (C/EBPs), *sterol regulatory binding protein 1* (SREBP1), PPARγ and EBFs in adipocyte development. Interestingly, we also identify a C/EBP motif, V.CEBPDELTA.Q6 at the 2 hour and 4 hour time-point (not shown). According to unpublished Chip-chip data performed in our lab, both C/EBPα and C/EBPβ are transcriptional targets of CREB and knockdown of C/EBPα and C/EBPβ significantly reduced neurite outgrowth *in vitro*. Another interesting result is the high ranking of the TRANSFAC motif V.TST1.01 in Table 4.7.

Predictor	tVIM	tVIM rank	pVIM	pVIM rank
V.E2F.Q6.01	-0.158	1	-0.129	1
V.TST1.01	-0.129	2	-0.01	22
V.MYB.Q3	-0.12	3	-0.038	3
V.LRF.Q2	0.096	4	0	NA
V.AP1.Q4.01	0.067	5	0.02	9
V.E2A.Q2	0.063	6	0.015	12
V.GEN.INI3.B	0.061	7	0.015	17
V.PAX4.03	0.056	8	0.053	2
V.OCT1.Q5.01	-0.055	9	-0.009	26
V.EBF.Q6	0.055	10	0.036	4
V.SP3.Q3	0.053	11	0.031	5
V.PPARA.01	0.05	12	0.029	6
V.POU6F1.01	-0.05	13	-0.027	8
V.MRF2.01	-0.046	15	-0.02	10
V.CETS1P54.02	-0.041	18	-0.016	11
V.VDR.Q3	0.04	19	0.027	7

Table 4.7: Top ranked predictors by pVIM and tVIM. Response variable Y represents log-fold gene expression in cultured F11 cells at 48 hours after Forskolin stimulation with respect to control.

This motif is bound by the *suppressed cAMP-inducible POU protein* (Scip alias Tst-1). Gondré *et al.* [38] have studied the function of Scip in schwann cells, which are glia (non-neuronal cells) in the peripheral nervous system. The expression of Scip is required for the establishment of normal nerves and it is re-expressed during regeneration. Furthermore, regeneration and hypertrophy of axons and myelin is markedly accelerated in transgenic mice expression a ΔScip transgene [38]. Although the fact that we identify Tst-1 as an important regulator of neuronal F11 cells may be surprising, it may be interesting to further study the role of this TF in neurons. Interactions between neurons and glial cells play important roles in regulating key events of development and regeneration of the CNS. Also, Table 4.6 and 4.7 list another POU-domain motif, V.POU6F1.01. The various members of

the POU family have a wide variety of functions, all of which are related to the development of an organism.

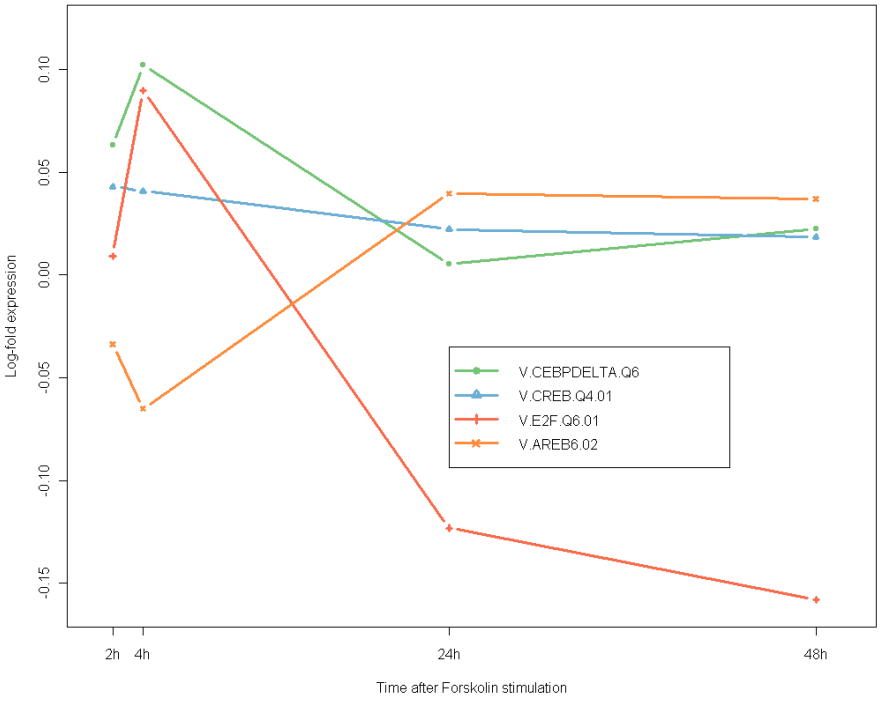

Figure 4.5: Plot of tVIM estimates versus time following Forskolin stimulation for several TRAP TF11 predictors associated to gene expression changes in F11 cells in response to Forkolin stimulation.

Altogether the results we present here identify several known and some putative novel DNA binding motifs that correspond to TFs which are likely to be important in the transcriptional regulatory network underlying neuronal regeneration. The VIM parameters allow us to estimate the variable importance for several highly ranked TRAP TF11 predictors at several time-points to get some insight into the dynamic activity of the corresponding TFs, as we did in the analysis of yeast cell cycle gene expression data in Section 4.4. The plot is drawn in Figure 4.5.

4.6 Discussion

The application of the variable importance estimation framework we consider in this chapter to analyze real experimental gene expression data provides biologically meaningful results. In particular, when genomewide time-course profiles of gene expression are available, it allows us to identify parameters that can be interpreted as representations of dynamic activity of transcriptional regulators underlying the observed patterns of gene expression. In Section 4.4 we validated the results we obtained on yeast cell cycle data against the literature on transcriptional regulation of the yeast cell cycle. This literature is largely based on analysis of *in vivo* binding data, such as ChIP-chip assays. Naturally, when such binding data is available, it can be used to replace or complement the surrogate predictors we use, i.e. `TRAP` predictors that represent binding affinities derived *in silico* according to a biophysical model of DNA binding by TFs. However, our main goal is to infer as much as possible about context specific regulatory effects of TFs on observed gene expression in absence of these data. The results of the analysis of yeast gene expression data in Section 4.4 show that the use of surrogate binding affinities as obtained using `TRAP` enables us to reconstruct the time dependent effect of several known cell cycle TFs such as `MBP1`, `MCM1`, `FKH2` and `SWI5` remarkably well.

In order to be able to use experimental *in vivo* binding data, these data should be obtained under the same experimental conditions under which the gene expression was measured. For mammalian gene expression experiments, such experimental binding data is typically only available for a couple of TFs or even completely lacking. This is also the case for the gene expression data from the *in vitro* biological model of neuronal regeneration we analyzed in Section 4.5. We identified known and putative novel TFs that are associated to patterns of early and late gene expression changes in F11 cells in response to Forskolin stimulation. The sign of the VIM parameter can be used to distinguish between transcriptional activators and repressors of gene expression. For instance, we estimated a positive value for the VIM of `V.E2F.Q6.01` at 2h following Forskolin stimulation and negative values, indicating repression of genes, for `V.E2F.Q6.01` at the two later time-points. Such information on dynamic activity of TFs (see Figure 4.5) is important for understanding the evolution of transcriptional regulatory networks in time.

Five

Conclusion

In this concluding chapter, we apply the methods discussed in the previous chapters to data from the neuronal injury model introduced in chapter 1 and build a network of transcription factors and their target genes involved in neuronal regeneration.

5.1 The transcriptional network underlying neuronal outgrowth

Condition-specific and time-dependent transcriptional GRNs (see Section 1.2) underlie the coordinated expression of genes involved in all biological processes. Insight into these networks is crucial for the understanding of biological systems under both normal and pathological conditions. In Chapters 2, 3 and 4 we have studied and developed different computational statistical methods that can be used to gain insight into different aspects of transcriptional networks through analysis of gene expression, DNA sequence and other relevant biological data. Here, we combine the different methods and integrate the most important findings into a transcriptional GRN that contains interactions between TFs and target genes involved in neuronal outgrowth. The network we build here is a directed network, with edges between source nodes and target nodes. The TFs we identified as putative regulators of neuronal outgrowth associated gene expression are the source nodes, whereas the genes that we predict to be transcriptionally regulated by these TFs are the corresponding targets. We base our network upon the F11 gene expression data introduced in Section 3.6. There we defined early and late Forskolin responsive genes and inferred transcriptional regulators that are associated to changes in gene expression in F11 cells in response to Forskolin. In Section 4.5 we estimated the relative importance of these TFs and distinguished between activators and repressors of neuronal outgrowth associated gene expression. Here, we build our GRN in the following way.

I Based on the results in Sections 3.6.1 and 4.5, we select 9 TFBS motifs (see Table 5.1) corresponding to some known and several putative novel regulators of neuronal outgrowth associated gene expression. We use these TFs as source nodes in our network. Temporally dynamic activity of TFs is believed to be important in transcriptional regulation. The results from our data analysis in Section 4.5 allow us to discern between early and late transcriptional activators and repressors. TFs for which the sign of the estimated VIM is positive are considered activators and repressors correspond to TFs with a negative VIM. TFs binding to the motifs in Table 5.1 with a VIM that ranks among the top 10 for either the 2 hour or 4 hour time-point are considered to be early regulators. Analogously, TFs binding to motifs that rank among the top 10 most important for the 24 hour or 48 hour time-point are putative late regulators.

II We use `LLM3D` to predict TF-target-gene relationships between the TFs in Table 5.1 and the sets of early and late responsive genes based upon the clustering described in Section 3.6.1. The resulting predictions determine the edges between TFs and target genes in our GRN.

Figure 5.1 contains a static image of the network that contains all interactions represented by gray edges. To visualize the dynamics of the network, we present two additional images of the same network that focus on the early and late gene expression changes. In these figures, edges between active regulators and their targets are colored. Green edges denote interactions between activators and their up-regulated targets, whereas red edges represent interactions between repressors and their down-regulated targets. The resulting networks

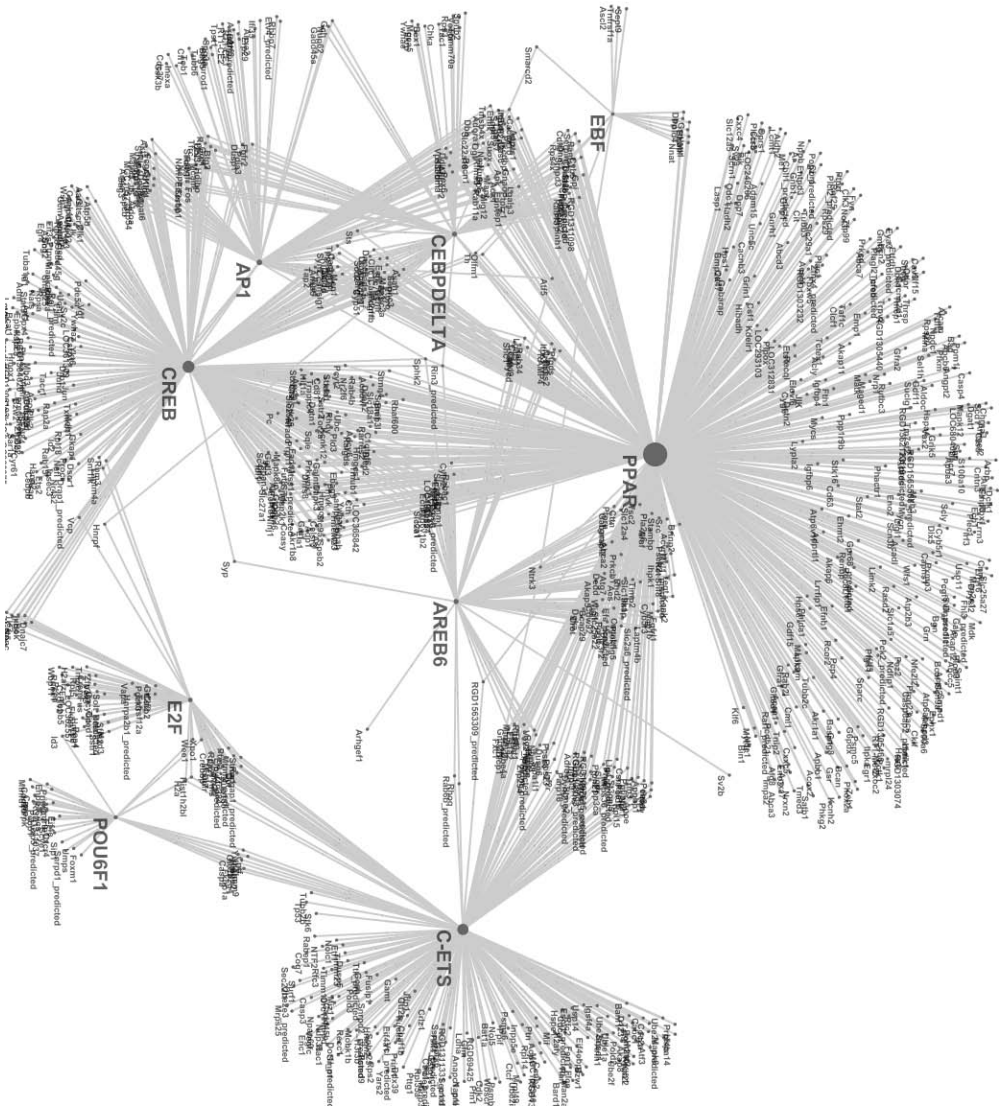

Figure 5.1: The GRN underlying neuronal outgrowth associated gene expression.

are visualized in Figures 5.2 and 5.3. Figure 5.2 highlights the early and Figure 5.3 the late interactions. These figures illustrate how outgrowth associated genes are predicted to be regulated by TFs across time.

The networks we present here suggest several hypotheses for further experimental validation. The implied presence of distinct early and late gene expression changes and the underlying dynamic activity of the TFs that regulate different targets at successive time-points are particularly interesting. With experimental techniques such as ChIP-chip or ChIP-seq, binding of TFs to the promoter regions of their predicted targets can be validated. However, experimental binding data are also noisy and occupation of a promoter by a TF does not necessarily imply a regulatory relationship. Therefore, to accurately define the most likely and biologically most important targets, statistical methods that integrate context-specific gene expression, DNA sequence *and* experimental binding data should be further developed, for instance using a Bayesian approach. Another interesting direction for future research could be to integrate data from related but different contexts. For instance, there is not one single biological model for neuronal regeneration. Separate analysis of gene expression data from *in vivo* and *in vitro* models of neuronal regeneration typically reveals different sets of regulated genes and, hence, different underlying transcriptional mechanisms. Biologically, one would expect significant overlap between related models, at least between the TFs involved. To characterize both the similarities and differences between the underlying networks presents the new challenge to develop computational statistical tools for modeling context-dependent changes in such biological networks.

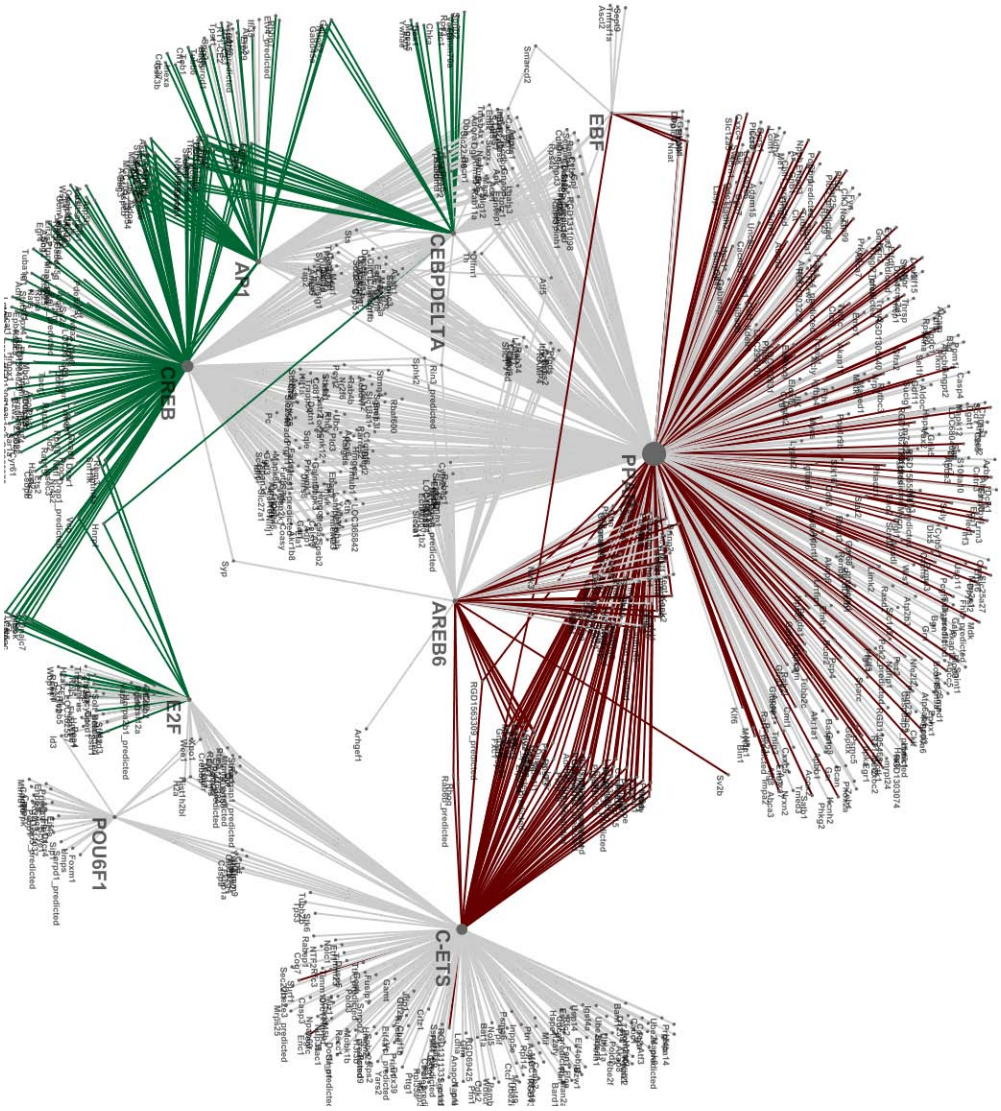

Figure 5.2: Same network as in Figure 5.1. Here, edges between early activators (green) and early repressors (red) and their predicted targets are highlighted.

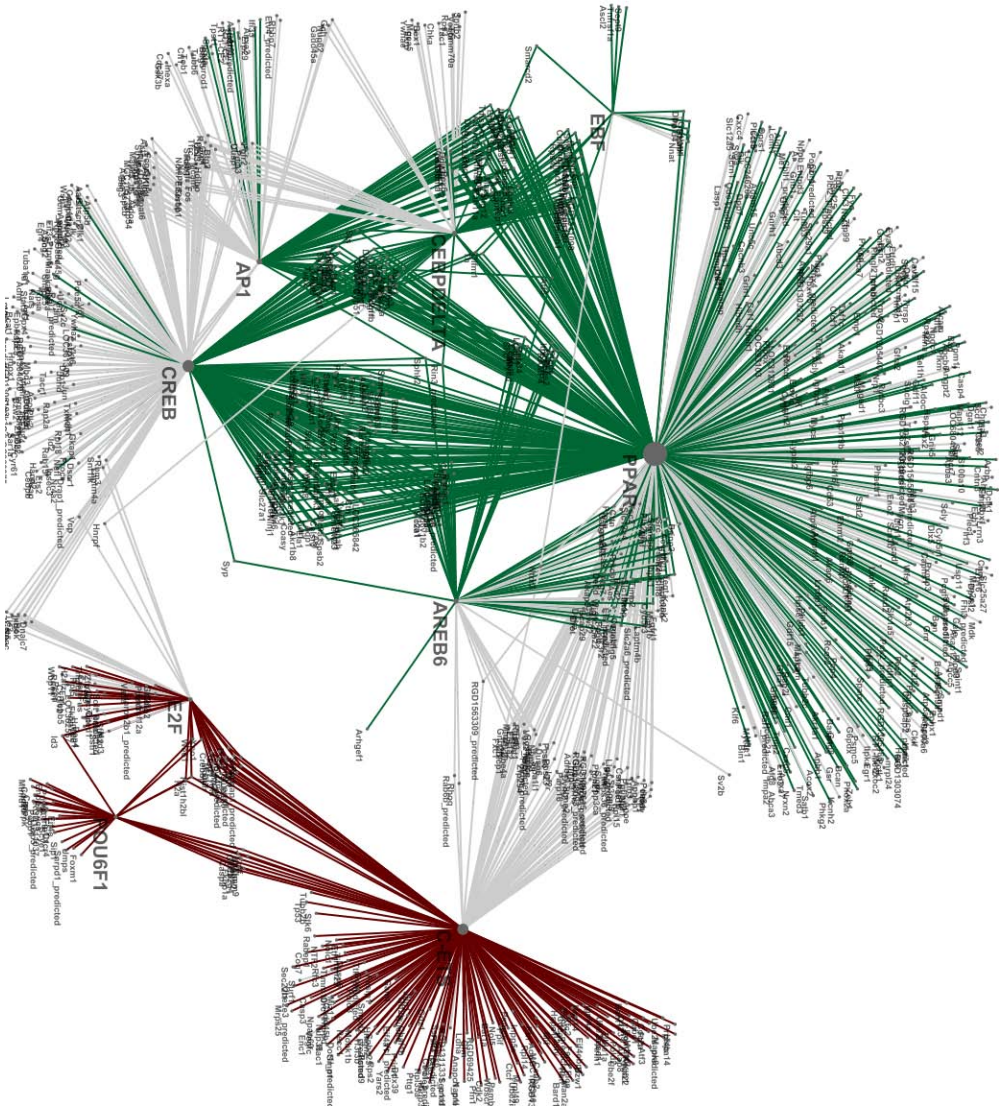

Figure 5.3: Same network as in Figure 5.1. Here, edges between late activators (green) and late repressors (red) and their predicted targets are highlighted.

TFBS ID	Motif logo	Time	Activity
V.AP1.Q4.01		Early	Activator
V.AP1.Q4.01		Late	Activator
V.AREB6.02		Early	Repressor
V.AREB6.02		Late	Activator
V.CREB.Q4.01		Early	Activator
V.CREB.Q4.01		Late	Activator
V.CEBPDELTA.Q6		Early	Activator
V.CEBPDELTA.Q6		Late	Activator
V.CETS1P54.02		Early	Repressor
V.CETS1P54.02		Late	Repressor
V.E2F.Q6.01		Early	Activator
V.E2F.Q6.01		Late	Repressor
V.EBF.Q6		Early	Repressor
V.EBF.Q6		Late	Activator
V.PPARA.01		Early	Repressor
V.PPARA.01		Late	Activator
V.POU6F1.01		Late	Repressor

Table 5.1: TFBS motif logos of TF source nodes of the GRN underlying neuronal outgrowth associated gene expression.

Appendix

A

LLM3D Supplementary Table

Supplementary Table 3. Predicted PPAR target genes with DR>SN expression and 'neuron differentiation' GO annotation, and predicted by both the non-conserved (rat only) and the HMR-conserved binding site options in LLM3D.

gene symbol	gene name	alternative gene symbol	cellular source	role in neurite outgrowth	supporting literature
Apbb1	amyloid beta (A4) precursor protein-binding, family B, member 1 (Fe65)	FE65	neuron	unknown	
Cdkn1c	cyclin-dependent kinase inhibitor 1C (P57)	p57; Kip2; p57KIP2; MGC112585	neuron	unknown	
Cnp1	2',3'-cyclic nucleotide 3' phosphodiesterase	CNPF; CNPI; CNPII; Cnp	neuron/glia	unknown	
Dapk3	death-associated protein kinase 3	Dapkl	neuron	unknown	
Dpysl3	dihydropyrimidinase-like 3	Crmp4; TUC-4b	neuron	inhibits	Alabed et al., 2007
Fgfr1	fibroblast growth factor receptor 1		neuron	stimulates	Hausott et al., 2008
Galr2	galanin receptor 2		neuron	stimulates	Hobson et al., 2006
Gnao	guanine nucleotide binding protein, alpha O	RATBPGTPC; Gnao1	neuron/glia	unknown	
Hes5	hairy and enhancer of split 5 (Drosophila)			inhibits	Sestan et al., 1999
L1cam	L1 cell adhesion molecule	Hyd; Hsas; NCAML1	neuron	stimulates	Panicker et al., 2003
Map1b	microtubule-associated protein 1B	Mtap1b	neuron/glia	inhibits	Bouquet et al., 2007
Nnat	neuronatin	Peg5; MGC156562	neuron	unknown	
Pgrmc1	progesterone receptor membrane component 1	MPR; 25Dx; VEMA; 25-Dx	neuron	unknown	
Pick1	protein interacting with PRKCA 1	Prkcabp	neuron	stimulates	Bartoe et al., 2006
Ptpn11	protein tyrosine phosphatase, non-receptor type 11	Shp2	neuron	stimulates	Rosario et al., 2007
Ret	ret proto-oncogene		neuron	stimulates	Luo et al., 2007
Rtn4	reticulon 4	r; Vp20; rat N; NI-250; MGC116054; rat NogoA	glia	inhibits	GrandPre et al., 2000; Chen et al., 2000
Rtn4rl2	reticulon 4 receptor-like 2	Ngrh1	neuron	inhibits	Venkatesh et al., 2005
S100a6	S100 calcium binding protein A6		glia	unknown	
Slit1	slit homolog 1 (Drosophila)	MEGF4	glia	inhibits	Brose et al., 1999
Thbs4	thrombospondin 4		neuron	stimulates	Arber and Caroni, 1995
Zfxh3	zinc finger homeobox 3		neuron	unknown	

Alabed YZ, Pool M, Ong Tone S, Fournier AE (2007) Identification of CRMP4 as a convergent regulator of axon outgrowth inhibition. J Neurosci 27:1702-1711

Arber S, Caroni P (1995) Thrombospondin-4, an extracellular matrix protein expressed in the developing and adult nervous system promotes neurite outgrowth. J Cell Biol 131:1083-1094

Bartoe JL, McKenna WL, Quan TK, Stafford BK, Moore JA, Xia J, Takamiya K, Huganir RL, Hinck L (2006) Protein interacting with C-kinase 1/protein kinase Calpha-mediated endocytosis converts netrin-1-mediated repulsion to attraction. J Neurosci 26:3192-3205

Bouquet C, Ravaille-Veron M, Propst F, Nothias F (2007) MAP1B coordinates microtubule and actin filament remodeling in adult mouse Schwann cell tips and DRG neuron growth cones. Mol Cell Neurosci 36:235-247

Brose K, Bland KS, Wang KH, Arnott D, Henzel W, Goodman CS, Tessier-Lavigne M, Kidd T (1999) Slit proteins bind Robo receptors and have an evolutionarily conserved role in repulsive axon guidance. Cell 96:795-806

Chen MS, Huber AB, van der Haar ME, Frank M, Schnell L, Spillmann AA, Christ F, Schwab ME (2000) Nogo-A is a myelin-associated neurite outgrowth inhibitor and an antigen for monoclonal antibody IN-1. Nature 403:434-439

GrandPré T, Nakamura F, Vartanian T, Strittmatter SM (2000) Identification of the Nogo inhibitor of axon regeneration as a Reticulon protein. Nature 403:439-444

Hausott B, Schlick B, Vallant N, Dorn R, Klimaschewski L (2008) Promotion of neurite outgrowth by fibroblast growth factor receptor 1 overexpression and lysosomal inhibition of receptor degradation in pheochromocytoma cells and adult sensory neurons. Neuroscience 153:461-473

Hobson SA, Holmes FE, Kerr NC, Pope RJ, Wynick D (2006) Mice deficient for galanin receptor 2 have decreased neurite outgrowth from adult sensory neurons and impaired pain-like behaviour. J Neurochem 99:1000-1010

Luo W, Wickramasinghe SR, Savitt JM, Griffin JW, Dawson TM, Ginty DD (2007) A hierarchical NGF signaling cascade controls Ret-dependent and Ret-independent events during development of nonpeptidergic DRG neurons. Neuron 54:739-754

Panicker AK, Buhusi M, Thelen K, Maness PF (2003) Cellular signalling mechanisms of neural cell adhesion molecules. Front Biosci 8:900-911

Rosário M, Franke R, Bednarski C, Birchmeier W (2007) The neurite outgrowth multiadaptor RhoGAP, NOMA-GAP, regulates neurite extension through SHP2 and Cdc42. J Cell Biol 178:503-516

Sestan N, Artavanis-Tsakonas S, Rakic P (1999) Contact-dependent inhibition of cortical neurite growth mediated by notch signaling. Science 286:741-746

Venkatesh K, Chivatakarn O, Lee H, Joshi PS, Kantor DB, Newman BA, Mage R, Rader C, Giger RJ (2005) The Nogo-66 receptor homolog NgR2 is a sialic acid-dependent receptor selective for myelin-associated glycoprotein. J Neurosci 25:808-822

B

LLM3D Package Description

How to use the `llm3d` package

Geert Geeven and Harold MacGillavry

January 28, 2010

1 Introduction

The `llm3d` package can be used to predict transcriptional regulators of functionally homogeneous and condition-specific sets of target genes from genome-wide expression data. `llm3d` simultaneously uses gene expression data, Gene Ontology (GO) and TFBS annotation in a combined statistical analysis that is based on log-linear models. The `llm3d` package can analyze any user-defined set of gene expression clusters (human, mouse, rat or yeast), but also includes two ready-to-use expression cluster sets (yeast and rat) that are discussed in the LLM3D paper by Geeven *et al.* (2010) [1]. This documentation explains how to analyze the example yeast gene expression data set using the `llm3d` package. An alternative approach to study TFBS enrichment is to predefine multiple sets of co-expressed genes sharing the same GO, and subsequently test each gene set for TFBS enrichment. This type of analysis, which we refer to as the multi gene sets by intersection (MGSI) approach (see Geeven *et al.* (2010) [1] for details), can also be performed by the `llm3d` package (see Section 7).

The `llm3d` package runs within the `R` environment for statistical computing and requires the following additional packages; `GO.db`, `MASS`, `annotate` and `gplots`. Make sure these are all installed before installing `llm3d`. The binary package for Windows was built under `R` version `2.10.1`. If you get any warnings when `llm3d` is loaded, make sure that you installed `R` $\geq$ `2.10.1` and also the latest versions of the packages that `llm3d` requires. To load the `llm3d` package in R use

```
> library(llm3d)
> data(llm3d.data)
```

Note that command lines can be copied and pasted in R. Commands must be copied one-by-one, selecting the command text line only and ignoring the R prompt ('>').

2 Defining the gene expression clusters

Here we consider a yeast gene expression data set published by Tu *et al.* (2005) on the regulation of yeast metabolic cycle genes. This study defined three large expression clusters of tightly co-regulated genes: an oxidative (Ox) cluster, a reductive building (Rb) cluster, and a reductive charging (Rc) cluster. To load the yeast dataset

```
> data(yeast.mc)
```

The `user.clusters` object is a list containing yeast Gene IDs for each of the defined gene-expression clusters. To see the cluster information included for the yeast dataset

```
> summary(user.clusters)
```

```
     Length Class  Mode
tu.Ox  988   -none- character
tu.Rb  924   -none- character
tu.Rc 1342   -none- character
```

Here, the variable `Length` indicates the number of gene IDs included in each of the clusters. Similarly, we have also included the rat DRG gene expression dataset described in Geeven *et al.* To load this dataset, use

```
> data(rat.drg)
```

To generate your own cluster object, first create tab-delimited `.txt` files containing your cluster gene IDs. Make one `.txt` file for each cluster. Note that `llm3d` only accepts Entrez Gene IDs (mouse, rat and human) or ORF IDs (yeast). Read the cluster `.txt` files by

```
> cluster1 <- read.delim("file1.txt", header=FALSE, stringsAsFactors=FALSE)
> cluster1 <- cluster1$V1
> cluster2 <- read.delim("file2.txt", header=FALSE, stringsAsFactors=FALSE)
> cluster2 <- cluster2$V1
```

Do this for each cluster you want to include. Create the new `user.clusters` object by

```
> user.clusters <- list(cluster1=cluster1, cluster2=cluster2)
```

Note that `llm3d` requires a named list-object like `user.clusters` in which the list elements are different character vectors containing unique Gene IDs. To see the cluster information of your newly generated cluster object

```
> summary(user.clusters)
```

Make sure each Gene ID occurs only once and that there is no overlap in Gene IDs between clusters. To check for overlap in your clusters

```
> K <- length(user.clusters)
> overlapping.IDs <- unique(unlist(user.clusters)[duplicated(unlist(user.clusters))])
> overlapping.IDs
```

To remove overlapping IDs from `user.clusters`

```
> for(i in 1:K) {user.clusters[[i]] <- setdiff(user.clusters[[i]],overlapping.IDs)}
```

3 TFBS and GO annotation data in LLM3D

The `llm3d` package includes TFBS and GO annotation for four different genomes. For human, mouse and rat TFBS annotation we used the position weight matrices available from TRANSFAC Professional, release 11.1 and scored all genes for the presence of TFBSs using the TRANSFAC MATCH-tool (see Geeven *et al.* for details). We also included annotation of human-mouse-rat conserved TFBSs, which is used as default when data from human, mouse or rat data are used as input. For the yeast TFBS annotation we used the Motifscanner tool with motifs from the improved regulatory map published by MacIsaac *et al.* (see Geeven *et al.* for details). For GO annotations we used GO Biological Process annotations from the GO database (`http://www.geneontology.org`). The TFBS and GO annotation data is stored in `llm3d.species.data`. This variable is a named list. The names correspond to the species that are available in `llm3d`

```
> names(llm3d.species.data)

[1] "yeast" "human" "mouse" "rat"
```

The variable `llm3d.species.data` contains the following data for yeast

```
> summary(llm3d.species.data$yeast)

                Length Class  Mode
llm3d.known.ids     2   -none- list
llm3d.tfbs.data     2   -none- list
llm3d.GO.targets 2675   -none- list
llm3d.GO.BP.ID   2675   -none- character
llm3d.GO.BP.term 2675   -none- character
llm3d.gene.names 5901   -none- character
```

The variable `llm3d.species.data` contains the following data for yeast

- `llm3d.known.ids` a named list of length 2 where each element is a vector of yeast gene IDs known to `llm3d` for both sets, i.e. `"single.species"` and `"conserved"`, of TFBS annotations (note that for yeast, these contain exactly the same data).
- `llm3d.tfbs.data` a named list of length 2 where each element is a matrix of yeast gene IDs (rows) and TFBSs (columns) containing the TFBS annotation of yeast genes for both TFBS annotations (note that for yeast, these contain exactly the same data).
- `llm3d.GO.targets` a list containing the yeast gene IDs annotated to GO Biological Process terms.
- `llm3d.GO.BP.ID` a vector of yeast GO Biological Process IDs known to `llm3d`.
- `llm3d.GO.BP.term` a vector of yeast GO Biological Process terms that are known to `llm3d`.

- `llm3d.gene.names` a vector of yeast gene names that are known to `llm3d`.

To check which yeast TFBSs are available

```
> get.TF.IDs("yeast")
```

The variable `llm3d.species.data$yeast$llm3d.tfbs.data $single.species` contains information about the presence or absence of TFBSs in yeast genes. To check presence of the first ten TFBSs in the first five yeast genes, use

```
> llm3d.species.data$yeast$llm3d.tfbs.data$single.species[1:5, 1:10]
```

```
        ABF1 ACE2 ADR1 AFT2 ARG80 ARG81 ARO80 ARR1 ASH1 AZF1
YAL001C    0    0    0    0     0     1     0    1    0    1
YAL002W    0    0    0    0     0     0     0    0    0    0
YAL003W    1    0    0    0     0     0     0    1    0    1
YAL005C    1    0    1    0     0     0     0    0    0    1
YAL007C    0    0    0    0     0     0     0    0    0    0
```

GO data for yeast are stored in `llm3d.species.data$yeast$llm3d.GO.targets`. For example

```
> head(llm3d.species.data$yeast$llm3d.GO.targets)
```

This latter command will show a list of the first five yeast GO terms and their associated yeast target gene IDs.

4 Selecting informative GO terms

The GO graph is a directed acyclic graph and the edges in the graph determine relationships between GO terms. This creates complicated dependencies between related GO terms when separately testing the associations of all GO terms with gene expression clusters, resulting in redundant lists of associated GO terms. High level terms in the GO graph, such as the root "biological process" or "biological regulation" are not particularly useful, since they are too general and cover many distinct processes. In contrast, the low level terms far away from the root, such as e.g. "nuclear mRNA 3'-splice site recognition", are very specific and contain only a handful of genes annotated to the term. Therefore, `llm3d` defines informative GO terms as a function of a threshold level (y) to select a representative set of GO terms for analysis. Pre-selecting informative GO terms helps to speed up the `llm3d` analysis and to reduce redundancy in the results. For a cluster x, Informative GO terms at level y are those terms that contain y associated genes in x, while all of its child terms contain $< y$ genes in x. By default, informative GO terms are selected at level 20 ($y = 20$). This means that all selected GO terms have at least 20 associated genes in at least one of the input clusters, while none of the selected GO terms will have child terms that are associated with more than 20 genes. You can change the y value by changing the `inf.GO` settings when calling the `llm3d` function (see item 5). Note that the selection of informative GO terms is based on the input gene clusters. Thus, small clusters require small y values, whereas the analysis of large clusters may benefit from larger y values. As a guideline you may want to use y values of about 5-10% of the size of the smallest input cluster.

5 Running LLM3D

`llm3d` uses log-linear modeling to test the association of user-provided gene clusters with the above described TFBS and GO annotation tables. Nine different dependency models can be tested by `llm3d`, however, only models M(7) and M(s) predict either pair-wise or complete dependencies between all three input variables (gene cluster, TFBS presence and GO annotation) and are reported by default (see Geeven *et al.* for details). Inference of transcriptional regulation is based on the significant association between gene cluster, TFBS presence and GO found by `llm3d`. To analyze the example yeast data set using the default settings

```
> llm3d.out <- llm3d("yeast", user.clusters=user.clusters,
user.ref=user.ref)
```

Note that as a consequence of fitting nine different models simultaneously, which require iterative methods for parameter estimation, for a large number of TFBS/GO pairs, the `llm3d` analysis typically takes several hours of run-time to complete depending on computer hardware. The arguments to the main function of `llm3d` are

```
> llm3d.out <- llm3d(user.species="yeast", user.clusters=user.clusters,
user.tfbs="conserved", user.ref="full.genome", user.GO=NULL,
user.TF=NULL, inf.GO=20, models=c("M7","MS"), mt.proc="BH", mt.q=0.1)
```

- `user.species` : the species to be used.
- `user.clusters` : a named list containing the clusters to be used (see item 2).
- `user.tfbs` : TFBS annotation to be used (see item 3).
- `user.ref` : the background gene set (see item 2).
- `inf.GO` : the y value to be used for the selection of informative GO terms (see item 4).
- `models` : the dependency models to be considered (see LLM3D paper by Geeven *et al.* (2010) [1]).
- `mt.proc` : the multiple testing procedure to be applied (default is BH, Benjamini Hochberg).
- `mt.q` : the multiple testing q-value cut-off to be applied.

To consult the help-page of the `llm3d` function

```
> help(llm3d)
```

Any of the default settings can be changed. For instance, to perform the same analysis using a y value of 50 to select informative GO terms

```
> llm3d.out <- llm3d("yeast", user.clusters=user.clcusters,
user.ref=user.ref, inf.GO=50)
```

For human, mouse and rat data, the default setting `user.tfbs="conserved"` setting can be changed into `user.tfbs="single.species"` to include all TFBSs, instead of only human-mouse-reat conserved TFBSs. Note, that when `user.species="yeast"`, this setting is ignored and the yeast TFBS annotation is used. The `llm3d.out` object is a list containing the results of the `llm3d` analysis and also contains some user-specified arguments.

```
> summary(llm3d.out)
              Length Class      Mode
user.species  1      -none-     character
user.clusters 3      -none-     list
user.tfbs     1      -none-     character
user.ref      1      -none-     character
user.method   1      -none-     character
llm3d.results 9      data.frame list
```

The `llm3d.results` data frame lists the significant TFBS-GO pairs associated with the user-defined expression clusters together with the best-fitting `llm3d` model. You can write the results table to a tab-delimited `.txt` file for further inspection by

```
> write.table(llm3d.out$llm3d.results,
file="llm3d_results.txt", sep="\t", row.names=FALSE, quote=FALSE)
```

The `llm3d_results.txt` file now contains a tab-separated table with the following information

- `TF.ID`: transcription factor binding site ID.
- `GO.ID`: GO biological process ID.
- `llm3d.best.model`: fitted `llm3d` model with lowest observed AIC for TFBS/GO pair (note that only models specified by the user in the argument `models` are listed).
- `llm3d.H0.pvalue`: p-value corresponding to testing H0 (complete independence of TFBS/GO/expression for TFBS/GO pair).
- `corrected.pvalue`: p-value corrected for multiple testing with `mt.proc`.
- `REF.residual`: the observed enrichment of target genes in the background gene set.
- `tu.Ox.residual`: the observed enrichment of target genes in cluster `tu.Ox`.
- `tu.Rb.residual`: the observed enrichment of target genes in cluster `tu.Rb`.
- `tu.Rc.residual`: the observed enrichment of target genes in cluster `tu.Rc`.

Note that target genes are defined here as genes with a binding site for the `TF.ID` AND annotated with the `GO.ID` under consideration.

6 Visualizing LLM3D results

To visualize the data use

```
> plot.llm3d(llm3d.out, compare.clust=c(1, 2), n.top.tf=20,
n.top.GO=20)
```

Instead of simply ranking associated TFBS-GO pairs according to their observed p-values, `llm3d` ranks all TFBS and GO terms based on how well their enrichment discriminates between the clusters. For this, the residual scores reported in `the llm3d_results.txt` file are used. The `plot.llm3d` function function will generate a heatmap that shows how the 20 most discriminative TFBSs and the 20 most discriminative GO terms are associated with two of the user-specified clusters (e.g. cluster 1 and cluster 2). In the resulting heatmap rows represent TFBSs and columns represent GO IDs. The colors represent a relative enrichment score within the first cluster (red) with respect to the second cluster (green). For details on association scores and the selection of the most discriminative TFBSs and GO terms, see Geeven *et al.* Enrichment scores are calculated with respect to those clusters that are specified in `compare.clust`. By default the first two clusters are chosen, but this can be adjusted. To visualize the second and third cluster use

```
> plot.llm3d(llm3d.out, compare.clust=c(2, 3), n.top.tf=20,
n.top.GO=20)
```

Increasing the values for `n.top.tf` and `n.top.GO` will increase the number of TFBS/GO associations plotted. Note that R scales the size of the plot to fit the size of your computer screen. This will render the axis labels for large plots illegible. Therefore, the scoring matrix can also be saved as a tab-delimited `.txt` file for use in other visualization software (for instance `TIGR Multiexperiment Viewer`, which can be download from `http://www.tm4.org/mev.html`). To save the heatmap matrix for 50 TFBS and 30 GO terms

```
> heatmap <- plot.llm3d(llm3d.out, compare.clust=c(1,2),
n.top.tf=50, n.top.GO=30, return.matrix=TRUE)

> write.table(heatmap, file="heatmap.txt", sep="\t",
quote=FALSE, col.names=NA)
```

To list GO terms instead of GO IDs use

```
> heatmap.GO.terms <- get.GO.terms(user.species="yeast",
colnames(heatmap))
> colnames(heatmap) <- heatmap.GO.terms
> write.table(heatmap, file="heatmap_terms.txt", sep="\t",
quote=FALSE, col.names=NA)
```

For any TFBS/GO pair of interest, the associated Gene IDs in each of the expression clusters can be retrieved using for instance

```
> targets <- get.targets(user.species="yeast",
in.tf="CAD1", in.GO="GO:0006807", user.clusters=user.clusters)
> targets
```

7 Running MGSI

To compare `llm3d` with the MGSI approach, run this

```
> mgsi.out <- mgsi("yeast", user.clusters)
```

```
> write.table(mgsi.out$mgsi.results,
file="mgsi_results.txt", sep="\t", row.names=FALSE, quote=FALSE)
```

The mgsi_results.txt file now contains a table with the following information

- `TF`: transcription factor binding site ID.
- `GO.ID`: gene ontology biological process ID.
- `GO.term`: gene ontology biological process term.
- `mgsi.cluster`: input gene cluster with most enrichment (lowest p-value).
- `tu.Ox.mgsi.corrected.p`: enrichment p-value (based on Fisher's exact test) corrected for multiple testing using mt.proc for cluster tu.Ox.
- `tu.Rb.mgsi.corrected.p`: enrichment p-value (based on Fisher's exact test) corrected for multiple testing using mt.proc for cluster tu.Rb.
- `tu.Rc.mgsi.corrected.p`: enrichment p-value (based on Fisher's exact test) corrected for multiple testing using mt.proc for cluster tu.Rc.

Default settings for `mgsi` are the same as for `llm3d` and can be adjusted in the same way. `mgsi` performs a conventional 2-way enrichment analysis using Fisher's exact test. Hence, there is no need to select any of the 3-way dependency models as for `llm3d`. As a result, computational complexity of `mgsi` is much lower resulting in considerably lower run times.

References

[1] Geeven, G., MacGillavry, H., Eggers, R., Sassen, M., Verhaagen, J., Smit, A., De Gunst, M., and Van Kesteren, R. LLM3D: a log-linear modeling-based method to predict functional gene regulatory interactions from genome-wide expression data. *Submitted* (2010).

Acknowledgements

I am enormously grateful for the help and moral support I received from many people without whom this thesis could not have been written. First of all I thank my main supervisor Mathisca for her patience and guidance during the past four years. Your encouragement, experience with applied statistical methods and your efforts to keep me focused and move forward, while still allowing me plenty of freedom are at the basis of this work. The collaboration with Ronald and Harold from the department of MCN at the CNCR has been fundamental to this work. Ronald, your enthousiasm for and devotion to my project, your wide knowledge and positive attitude have been inspiring. It was a great pleasure discussing with you and Harold the biological side of the story. Also the numerous discussions with Guus have been invaluable. Joost and other people at the NIN I thank for their interest in my work and useful insights and feedback. I am indebted to all the members of the reading committee for their labor and helpful comments.

Mark van der Laan provided me with the great opportunity to spend four months at Berkeley to study his work on estimation of variable importance. Mark, the time and effort you devoted to familiarizing me to this subject are very much appreciated. All present and former colleagues and members of the Statistics with Life Sciences group at the VU I will remember for the pleasant and stimulating working environment and group meetings. In particular, warm greetings go to my roommates Willem, Ismaël, Rik, Rikkert, Wessel and Shota.

Joop, you provided me with a warm and cosy home when I was most in need. There never was a lack of Mad Sauce, beer or talentful students from the Royal Conservatory to join us for some cultural activities. I hope one day we can move back in together. Jeroen, you are both an everlasting true friend *and* relative and for that reason deserve to be mentioned twice. Thanks for always being there for me and for all that we have been through together the last 26 years. Roelande I particularly thank for sharing the frustration of early onset baldness and for the countless "loepzuivere" two-voice harmonies we sang about a grubby Eindhoven-based substandard food establishment. Final thanks go to my closest family, most notably my parents, who provided me with much more than I can mention here and always supported me through my scientific endeavours.

Thanks to all of you. Take care !

Samenvatting (Dutch Summary)

Sinds het einde van de vorige eeuw is de systematische studie van biologische netwerken sterk in opkomst. Ontwikkelingen in de moleculaire biologie hebben geleid tot vele nieuwe experimentele technieken waarmee biologische systemen in meer en meer detail kunnen worden onderzocht. Hedendaagse experimenten genereren grote hoeveelheden data. Het analyseren van deze data met als doel meer inzicht te krijgen in de onderliggende biologische processen, vereist de ontwikkeling van nieuwe concreet toepasbare statistische methoden.

In dit proefschrift ontwikkelen we computationele en statistische methoden die gebruikt kunnen worden om interacties tussen transcriptiefactoren en target genen te identificeren. Transcriptiefactoren zijn moleculen die aan het DNA in de kern van een cel binden en daardoor expressie van genen kunnen beïnvloeden. Het DNA in de celkern bevat, verspreid over chromosomen, vele verschillende genen. Ieder gen bevat een unieke code die bepaalt hoe een specifiek eiwit opgebouwd moet worden. Eiwitten zijn van essentieel belang voor de opbouw en het functioneren van cellen. Omdat in verschillende typen cellen verschillende eiwitten moeten worden aangemaakt, hebben cellen mechanismen om naar behoefte bepaalde genen te activeren of te deactiveren. Een van deze mechanismen is transcriptionele regulatie van genexpressie. De zogenaamde transcriptionele interacties tussen DNA bindende factoren en de target genen die ze activeren (tot expressie brengen), of juist deactiveren, zijn sterk afhankelijk van cellulaire condities en tijd en vormen een complex dynamisch transcriptioneel netwerk. Inzicht in deze netwerken is van fundamenteel belang om te begrijpen hoe cellen de expressie van genen reguleren en kunnen reageren op veranderende omstandigheden.

Het eerste hoofdstuk van dit proefschrift is een inleiding waarin kort de relevante biologische theorie over transcriptionele regulatie en genexpressie wordt behandeld. We bespreken de verschillende typen data die in dit proefschrift worden geanalyseerd en we beschouwen een voorbeeld van een genetisch netwerk. Verder introduceren we twee biologische modellen voor zenuwregeneratie die in dit proefschrift een speciale rol spelen. Een belangrijk doel van het onderzoek dat aan dit proefschrift ten grondslag ligt is namelijk om door analyse van data verkregen uit zenuwregeneratie-onderzoek inzicht te verschaffen in de netwerken van transcriptiefactoren en genen die betrokken zijn bij dit proces.

In hoofdstuk 2 beschrijven we een nieuwe computationele methode, genaamd `LLM3D`, die gebruikt kan worden om functionele bindings-plaatsen (de sequenties in het DNA waar transcriptiefactoren aan binden) te identificeren op basis van significante verrijking van deze sites in groepen functioneel homogene en co-gereguleerde genen. `LLM3D` fit zogenaamde

log-lineaire modellen voor drie-dimensionale kruistabellen. Voorspellingen van LLM3D zijn gebaseerd op afhankelijkheden tussen genexpressie, genfunctie en het voorkomen van bindings-plaatsen die door deze modellen worden geïmpliceerd. We valideren onze methode aan de hand van gistdata en laten zien dat transcriptionele targets met LLM3D nauwkeuriger worden voorspeld dan met een bestaande methode. Door LLM3D toe te passen op data uit het model voor zenuwregeneratie zijn we in staat een nieuwe transcriptiefactor te identificeren die een significante invloed heeft op de regeneratie van zenuwuitlopers.

In het derde hoofdstuk onderzoeken we hoe regressiemodellen kunnen worden gebruikt om variatie in genexpressie te modeleren als functie van het binden van transcriptiefactoren aan het DNA rondom de genen. We beschrijven een nieuwe strategie, genaamd GEMULA, die is gebaseerd op lineaire modellen. Dit leidt tot een snelle regressiemethode waarmee een breed scala aan plausibele en biologisch interpreteerbare modellen kan worden gefit. Deze modellen bevatten vaak vele variabelen (de transcriptiefactoren) en interacties tussen de variabelen, waardoor het relatieve effect van de afzonderlijke variabelen op de genexpressie variatie niet meteen duidelijk is. Daarom beschouwen we in hoofdstuk 4 een parameter die dit effect voor de afzonderlijke variabelen voorstelt. We bestuderen verschillende schatters voor deze parameter en laten aan de hand van een biologisch voorbeeld zien dat deze parameter een relevante interpretatie heeft.

Ter conclusie presenteren we in het laatste hoofdstuk een grafische voorstelling van het netwerk van transcriptiefactoren en genen dat betrokken is bij zenuwregeneratie. Dit netwerk is het resultaat van toepassing van de methoden uit de voorgaande hoofdstukken en bevat nieuwe, computationele voorspellingen van transcriptionele interacties die potentieel belangrijk zijn voor robuuste zenuwuitgroei. Het dient als startpunt voor verdere biologische karakterisering en validatie van de moleculaire mechanismen die kunnen leiden tot succesvolle zenuwregeneratie en is een voorbeeld van hoe toepassing van de in dit proefschrift ontwikkelde methoden kan leiden tot nieuwe biologische inzichten.

References

[1] Aerts, S., Thijs, G., Coessens, B., Staes, M., Moreau, Y., and De Moor, B. Toucan: deciphering the cis-regulatory logic of coregulated genes. *Nucleic Acids Res 31*, 6 (March 2003), 1753–1764.

[2] Akaike, H. Information Theory and an Extension of the Maximum Likelihood Principle. In *International Symposium on Information Theory, 2 nd, Tsahkadsor, Armenian SSR* (1973), pp. 267–281.

[3] Akaike, H. A new look at the statistical model identification. *Automatic Control, IEEE Transactions on 19*, 6 (1974), 716–723.

[4] Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., and Walter, P. *Molecular Biology of the Cell*, 5th ed. Garland Science, November 2008.

[5] Alon, U. Network motifs: theory and experimental approaches. *Nature Reviews Genetics 8*, 6 (June 2007), 450–461.

[6] Angelini, C., Cutillo, L., De Canditiis, D., Mutarelli, M., and Pensky, M. BATS: A Bayesian user-friendly software for Analyzing Time Series microarray experiments. *BMC Bioinformatics 9*, 415 (2008).

[7] Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., and Sherlock, G. Gene ontology: tool for the unification of biology. *Nature Genetics 25*, 1 (May 2000), 25–29.

[8] Bacon, A., Kerr, N. C. H., Holmes, F. E., Gaston, K., and Wynick, D. Characterization of an Enhancer Region of the Galanin Gene That Directs Expression to the Dorsal Root Ganglion and Confers Responsiveness to Axotomy. *J. Neurosci. 27*, 24 (2007), 6573–6580.

[9] Banerjee, N., and Zhang, M. Q. Identifying cooperativity among transcription factors controlling the cell cycle in yeast. *Nucl. Acids Res. 31*, 23 (2003), 7024–7031.

[10] Barabasi, A.-L., and Oltvai, Z. N. Network biology: understanding the cell's functional organization. *Nature Reviews Genetics 5*, 2 (February 2004), 101–113.

[11] BEER, M. A., AND TAVAZOIE, S. Predicting gene expression from sequence. *Cell 117*, 2 (Apr 2004), 185–98.

[12] BENJAMINI, Y., AND HOCHBERG, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological) 57*, 1 (1995), 289–300.

[13] BREIMAN, L. Random forests. *Machine Learning 45*, 1 (October 2001), 5–32.

[14] BREIMAN, L., FRIEDMAN, J., STONE, C. J., AND OLSHEN, R. A. *Classification and Regression Trees*. Chapman & Hall/CRC, January 1984.

[15] BRYNE, J. C., VALEN, E., TANG, M.-H. E., MARSTRAND, T., WINTHER, O., DA PIEDADE, I., KROGH, A., LENHARD, B., AND SANDELIN, A. Jaspar, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update. *Nucl. Acids Res. 36*, suppl_1 (January 2008), D102–106.

[16] BURNHAM, K., AND ANDERSON, D. *Model Selection and Multimodel Inference: a Practical Information Theoretic Approach*. Springer, New York, 2002.

[17] BUSSEMAKER, H., LI, H., AND SIGGIA, E. Regulatory element detection using correlation with expression. *NATURE GENETICS 27*, 2 (FEB 2001), 167–171.

[18] CHERRY, J. M., ADLER, C., BALL, C., CHERVITZ, S. A., DWIGHT, S. S., HESTER, E. T., JIA, Y., JUVIK, G., ROE, T., SCHROEDER, M., WENG, S., AND BOTSTEIN, D. Sgd: Saccharomyces genome database. *Nucleic acids research 26*, 1 (January 1998), 73–79.

[19] CHRISTENSEN, R. Log-linear models and logistic regression. In *Series in Statistics* (1997), Springer.

[20] COKUS, S., ROSE, S., HAYNOR, D., GRONBECH-JENSEN, N., AND PELLEGRINI, M. Modelling the network of cell cycle transcription factors in the yeast saccharomyces cerevisiae. *BMC Bioinformatics 7*, 1 (2006), 381.

[21] CONSORTIUM, G. O. The gene ontology (go) database and informatics resource. *Nucleic Acids Research 32*, suppl_1 (January 2004), D258–261.

[22] COSTIGAN, M., BEFORT, K., KARCHEWSKI, L., GRIFFIN, R., D'URSO, D., ALLCHORNE, A., SITARSKI, J., MANNION, J., PRATT, R., AND WOOLF, C. Replicate high-density rat genome oligonucleotide microarrays reveal hundreds of regulated genes in the dorsal root ganglion after peripheral nerve injury. *BMC Neuroscience 3*, 1 (2002), 16.

[23] DAS, D., BANERJEE, N., AND ZHANG, M. Q. Interacting models of cooperative gene regulation. *Proceedings of the National Academy of Sciences of the United States of America 101*, 46 (2004), 16234–16239.

[24] DAS, D., NAHLÉ, Z., AND ZHANG, M. Q. Adaptively inferring human transcriptional subnetworks. *Mol Syst Biol 2* (2006).

[25] DAS, D., PELLEGRINI, M., AND GRAY, J. W. A primer on regression methods for decoding *cis*-regulatory logic. *PLoS Comput Biol 5*, 1 (01 2009), e1000269.

[26] EFRON, B., HASTIE, T., JOHNSTONE, I., AND TIBSHIRANI, R. Least angle regression. *Annals of Statistics 32* (2002), 407–499.

[27] EFRON, B., HASTIE, T., JOHNSTONE, I., AND TIBSHIRANI, R. Least angle regression. *Annals of Statistics 32*, 2 (2004), 407–499.

[28] ELKON, R., LINHART, C., SHARAN, R., SHAMIR, R., AND SHILOH, Y. Genome-Wide In Silico Identification of Transcriptional Regulators Controlling the Cell Cycle in Human Cells. *Genome Research 13*, 5 (2003), 773–780.

[29] FORLUVOFT. `http://en.wikipedia.org/wiki/File:Gene2-plain.svg`.

[30] FRIEDMAN, J. Multivariate adaptive regression splines. *Annals of Statistics 19*, 1 (Mar 1991), 1–67.

[31] FRIEDMAN, N., LINIAL, M., NACHMAN, I., AND PE'ER, D. Using Bayesian networks to analyze expression data. *Journal of Computational Biology 7*, 3 (August 2000), 601–620.

[32] GAO, Y., HOU, J., BRYSON, J., BARCO, A., NIKULINA, E., SPENCER, T., MELLADO, W., KANDEL, E. R., AND FILBIN, M. T. Activated creb is sufficient to overcome inhibitors in myelin and promote spinal axon regeneration in vivo. *Neuron 44*, 4 (2004), 609–621.

[33] GARCIA-DOMINGUEZ, M., POQUET, C., GAREL, S., AND CHARNAY, P. Ebf gene function is required for coupling neuronal differentiation and cell cycle exit. *Development 130*, 24 (2003), 6013–6025.

[34] GAREL, S., MARÍN, F., MATTÉI, M., VESQUE, C., VINCENT, A., AND CHARNAY, P. Family of `Ebf/Olf-1`-related genes potentially involved in neuronal differentiation and regional specification in the central nervous system. *Developmental Dynamics 210*, 3 (1997), 191–205.

[35] GASCH, A. P., SPELLMAN, P. T., KAO, C. M., CARMEL-HAREL, O., EISEN, M. B., STORZ, G., BOTSTEIN, D., AND BROWN, P. O. Genomic expression programs in the response of yeast cells to environmental changes. *Molecular biology of the cell 11*, 12 (December 2000), 4241–4257.

[36] GENTLEMAN, R. C., CAREY, V. J., BATES, D. M., BOLSTAD, B., DETTLING, M., DUDOIT, S., ELLIS, B., GAUTIER, L., GE, Y., GENTRY, J., HORNIK, K., HOTHORN, T., HUBER, W., IACUS, S., IRIZARRY, R., LEISCH, F., LI, C., MAECHLER, M., ROSSINI, A. J., SAWITZKI, G., SMITH, C., SMYTH, G., TIERNEY, L., YANG, J. Y., AND ZHANG, J. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol 5*, 10 (2004).

[37] GERSTEIN, M. B., BRUCE, C., ROZOWSKY, J. S., ZHENG, D., DU, J., KORBEL, J. O., EMANUELSSON, O., ZHANG, Z. D., WEISSMAN, S., AND SNYDER, M. What is a gene, post-encode? history and updated definition. *Genome Research 17*, 6 (June 2007), 669–681.

[38] GONDRE, M., BURROLA, P., AND WEINSTEIN, D. E. Accelerated Nerve Regeneration Mediated by Schwann Cells Expressing a Mutant Form of the POU Protein SCIP. *J. Cell Biol. 141*, 2 (1998), 493–501.

[39] GOODRUM, J., WEAVER, J., GOINES, N., AND BOULDIN, T. W. Fatty acids from degenerating myelin lipids are conserved and reutilized for myelin synthesis during regeneration in peripheral nerve. *Journal of Neurochemistry 65*, 4 (1995), 1752–1759.

[40] HANNENHALLI, S. Eukaryotic transcription factor binding sites–modeling and integrative search methods. *Bioinformatics 24*, 11 (2008), 1325–1331.

[41] HARBISON, C. T., GORDON, D. B., LEE, T. I., RINALDI, N. J., MACISAAC, K. D., DANFORD, T. W., HANNETT, N. M., TAGNE, J.-B., REYNOLDS, D. B., YOO, J., JENNINGS, E. G., ZEITLINGER, J., POKHOLOK, D. K., KELLIS, M., ROLFE, P. A., TAKUSAGAWA, K. T., LANDER, E. S., GIFFORD, D. K., FRAENKEL, E., AND YOUNG, R. A. Transcriptional regulatory code of a eukaryotic genome. *Nature 431* (September 2004), 99 – 104.

[42] HARTEMINK, A. J. Reverse engineering gene regulatory networks. *Nat Biotechnol 23*, 5 (May 2005), 554–555.

[43] HAYETE, B., GARDNER, T., AND COLLINS, J. Size matters: network inference tackles the genome scale. *Mol Syst Biol 3* (2007).

[44] HECKER, M., LAMBECK, S., TOEPFER, S., VAN SOMEREN, E., AND GUTHKE, R. Gene regulatory network inference: Data integration in dynamic models–a review. *Biosystems 96*, 1 (2009), 86 – 103.

[45] HERDEGEN, T., KUMMER, W., FIALLOS, C., LEAH, J., AND BRAVO, R. Expression of c-jun, jun b and jun d proteins in rat nervous system following transection of vagus nerve and cervical sympathetic trunk. *Neuroscience 45*, 2 (1991), 413 – 422.

[46] HERDEGEN, T., SKENE, P., AND BÄHR, M. The c-jun transcription factor –bipotential mediator of neuronal death, survival and regeneration. *Trends in Neurosciences 20*, 5 (1997), 227–231.

[47] HIHI, A., MICHALIK, L., AND WAHLI, W. Ppars: transcriptional effectors of fatty acids and their derivatives. *Cellular and Molecular Life Sciences 59*, 5 (2002), 790–798.

[48] HIPPENMEYER, S., VRIESELING, E., SIGRIST, M., PORTMANN, T., LAENGLE, C., LADLE, D. R., AND ARBER, S. A developmental switch in the response of drg neurons to ets transcription factor signaling. *PLoS Biol 3*, 5 (04 2005), e159.

[49] HUANG, D. W., SHERMAN, B. T., AND LEMPICKI, R. A. Systematic and integrative analysis of large gene lists using david bioinformatics resources. *Nature Protocols 4*, 1 (December 2008), 44–57.

[50] HUANG, D. W., SHERMAN, B. T., AND LEMPICKI, R. A. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucl. Acids Res. 37*, 1 (January 2009), 1–13.

[51] JORDAN, M. *Learning in graphical models*. Kluwer Academic Publishers, 1998.

[52] JOTHI, R., BALAJI, S., WUSTER, A., GROCHOW, J. A., GSPONER, J., PRZYTYCKA, T. M., ARAVIND, L., AND BABU, M. M. Genomic analysis reveals a tight link between transcription factor dynamics and regulatory network architecture. *Molecular Systems Biology 5* (August 2009).

[53] JUNG, AND KIM, Y.-J. J. Intrinsic variability of gene expression encoded in nucleosome positioning sequences. *Nature genetics* (March 2009).

[54] KEL, A., GOSSLING, E., REUTER, I., CHEREMUSHKIN, E., KEL-MARGOULIS, O., AND WINGENDER, E. MATCH^{TM}: a tool for searching transcription factor binding sites in DNA sequences. *Nucl. Acids Res. 31*, 13 (2003), 3576–3579.

[55] KLIEWER, S. A., UMESONO, K., NOONAN, D. J., HEYMAN, R. A., AND EVANS, R. M. Convergence of 9-cis retinoic acid and peroxisome proliferator signalling pathways through heterodimer formation. *Nature 358*, 6389 (1992), 771–774.

[56] LAHIRI, P., Ed. *Model selection*, vol. 38 of *IMS Lecture Notes & Monographs Series*. Institute of Mathematical Statistics, 2001.

[57] LAURITZEN, S. *Graphical Models*. Oxford University Press, 1996.

[58] LEE, H. K., HSU, A. K., SAJDAK, J., QIN, J., AND PAVLIDIS, P. Coexpression Analysis of Human Genes Across Many Microarray Data Sets. *Genome Research 14*, 6 (2004), 1085–1094.

[59] LEE, T. I., RINALDI, N. J., ROBERT, F., ODOM, D. T., BAR-JOSEPH, Z., GERBER, G. K., HANNETT, N. M., HARBISON, C. T., THOMPSON, C. M., SIMON, I., ZEITLINGER, J., JENNINGS, E. G., MURRAY, H. L., GORDON, D. B., REN, B., WYRICK, J. J., TAGNE, J.-B., VOLKERT, T. L., FRAENKEL, E., GIFFORD, D. K., AND YOUNG, R. A. Transcriptional regulatory networks in saccharomyces cerevisiae. *Science 298*, 5594 (October 2002), 799–804.

[60] LEE, W.-P., AND TZOU, W.-S. Computational methods for discovering gene networks from expression data. *Brief Bioinform 10*, 4 (2009), 408–423.

[61] LEÓN, D. M., WELCHER, A., NAHIN, R., LIU, Y., RUDA, M., SHOOTER, E., AND C.A., M. Fatty acid binding protein is induced in neurons of the dorsal root ganglia after peripheral nerve injury. *Journal of Neuroscience Research 44*, 3 (1996), 283–292.

[62] LIBERG, D., SIGVARDSSON, M., AND AKERBLAD, P. The EBF/Olf/Collier Family of Transcription Factors: Regulators of Differentiation in Cells Originating from All Three Embryonal Germ Layers. *Mol. Cell. Biol. 22*, 24 (2002), 8389–8397.

[63] LIU, J.-W., ALMAGUEL, F. G., BU, L., DE LEON, D. D., AND DE LEON, M. Expression of e-fabp in pc12 cells increases neurite extension during differentiation: involvement of n-3 and n-6 fatty acids. *Journal of Neurochemistry 106*, 5 (2008), 2015–2029.

[64] LODISH, H., BERK, A., KAISER, C. A., KRIEGER, M., SCOTT, M. P., BRETSCHER, A., PLOEGH, H., AND MATSUDAIRA, P. *Molecular Cell Biology (Lodish, Molecular Cell Biology)*, 6th ed. W. H. Freeman, June 2007.

[65] MACGILLAVRY, H. D., STAM, F. J., SASSEN, M. M., KEGEL, L., HENDRIKS, W. T. J., VERHAAGEN, J., SMIT, A. B., AND VAN KESTEREN, R. E. NFIL3 and cAMP Response Element-Binding Protein Form a Transcriptional Feedforward Loop that Controls Neuronal Regeneration-Associated Gene Expression. *J. Neurosci. 29*, 49 (2009), 15542–15550.

[66] MACISAAC, K., WANG, T., GORDON, D. B., GIFFORD, D., STORMO, G., AND FRAENKEL, E. An improved map of conserved regulatory sites for saccharomyces cerevisiae. *BMC Bioinformatics 7*, 1 (March 2006), 113+.

[67] MADPRIME. `http://en.wikipedia.org/wiki/File:Genetic_code.svg`.

[68] MATYS, V., KEL-MARGOULIS, O. V., FRICKE, E., LIEBICH, I., LAND, S., BARRE-DIRRIE, A., REUTER, I., CHEKMENEV, D., KRULL, M., HORNISCHER, K., VOSS, N., STEGMAIER, P., LEWICKI-POTAPOV, B., SAXEL, H., KEL, A. E., AND WINGENDER, E. Transfac® and its module transcompel®: transcriptional gene regulation in eukaryotes. *Nucleic Acids Research 34* (2006), D108.

[69] MCTIGUE, D., TRIPATHI, R., WEI, P., AND A.T., L. The ppar gamma agonist pioglitazone improves anatomical and locomotor recovery after rodent spinal cord injury. *Experimental Neurology 205*, 2 (2007), 396–406.

[70] MCTIGUE, D. M. Potential therapeutic targets for pparγ after spinal cord injury. *PPAR Research 2008* (2008).

[71] MIN, I. M., PIETRAMAGGIORI, G., KIM, F. S., PASSEGUÉ, E., STEVENSON, K. E., AND WAGERS, A. J. The transcription factor egr1 controls both the proliferation and localization of hematopoietic stem cells. *Cell Stem Cell 2*, 4 (2008), 380–391.

[72] NAM, D., AND KIM, S.-Y. Gene-set approach for expression pattern analysis. *Brief Bioinform 9*, 3 (May 2008), 189–197.

[73] NISHII, R. Asymptotic properties of criteria for selection of variables in multiple regression. *The Annals of Statistics 12*, 2 (1984), 758–765.

[74] NOGALES-CADENAS, R., CARMONA-SAEZ, P., VAZQUEZ, M., VICENTE, C., YANG, X., TIRADO, F., CARAZO, J. M. M., AND PASCUAL-MONTANO, A. Genecodis: interpreting gene lists through enrichment analysis and integration of diverse biological information. *Nucleic acids research 37*, Web Server issue (July 2009), W317–322.

[75] OHNUMA, S.-I., AND HARRIS, W. A. Neurogenesis and the cell cycle. *Neuron 40*, 2 (2003), 199–208.

[76] PARK, S.-W., YI, J.-H., MIRANPURI, G., SATRIOTOMO, I., BOWEN, K., RESNICK, D. K., AND VEMUGANTI, R. Thiazolidinedione Class of Peroxisome Proliferator-Activated

Receptor Îş Agonists Prevents Neuronal Damage, Motor Dysfunction, Myelin Loss, Neuropathic Pain, and Inflammation after Spinal Cord Injury in Adult Rats. *Journal of Pharmacology and Experimental Therapeutics 320*, 3 (2007), 1002–1012.

[77] PHUONG, T. M., LEE, D., AND LEE, K. H. Regression trees for regulatory element identification. *Bioinformatics 20*, 5 (2004), 750–757.

[78] POKHOLOK, D. K., HARBISON, C. T., LEVINE, S., COLE, M., HANNETT, N. M., LEE, T. I., BELL, G. W., WALKER, K., ROLFE, P. A., HERBOLSHEIMER, E., ZEITLINGER, J., LEWITTER, F., GIFFORD, D. K., AND YOUNG, R. A. Genome-wide map of nucleosome acetylation and methylation in yeast. *Cell 122*, 4 (2005), 517 – 527.

[79] R. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, 2005.

[80] RAHMANN, S., MÜLLER, T., AND VINGRON, M. On the power of profiles for transcription factor binding site detection. *Statistical Applications in Genetics and Molecular Biology 2*, 1 (2007), 7.

[81] RAIVICH, G., M., B., DA COSTA, C., IWATA, O., ..., AGUZZI, A., WAGNER, E., AND A., B. The ap-1 transcription factor c-jun is required for efficient axonal regeneration. *Neuron 43*, 1 (2004), 57–67.

[82] REIMAND, J., KULL, M., PETERSON, H., HANSEN, J., AND VILO, J. g:profiler–a web-based toolset for functional profiling of gene lists from large-scale experiments. *Nucleic acids research 35*, Web Server issue (July 2007).

[83] RITCHIE, M. E., SILVER, J., OSHLACK, A., HOLMES, M., DIYAGAMA, D., HOLLOWAY, A., AND SMYTH, G. K. A comparison of background correction methods for two-colour microarrays. *Bioinformatics 23*, 20 (2007), 2700–2707.

[84] ROBSON, L., DYALL, S., SIDLOFF, D., AND MICHAEL-TITUS, A. Omega-3 polyunsaturated fatty acids increase the neurite outgrowth of rat sensory neurones throughout development and in aged animals. *Neurobiology of Aging* (2008).

[85] ROIDER, H. G., KANHERE, A., MANKE, T., AND VINGRON, M. Predicting transcription factor affinities to DNA from a biophysical model. *Bioinformatics 23*, 2 (2007), 134–141.

[86] SCHÄFER, J., AND STRIMMER, K. An empirical Bayes approach to inferring large-scale gene association networks. *Bioinformatics 21*, 6 (Mar 2005), 754–64.

[87] SCHÄFER, J., AND STRIMMER, K. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical Applications in Genetics and Molecular Biology 4*, 1 (2005), art.32.

[88] SCHMITT, A., BREUER, S., LIMAN, J., BUSS, A., SCHLANGEN, C., PECH, K., HOL, E., BROOK, G., NOTH, J., AND SCHWAIGER, F.-W. Identification of regeneration-associated genes after central and peripheral nerve injury in the adult rat. *BMC Neuroscience 4*, 1 (2003), 8.

[89] SCHNEIDER, T. D., AND STEPHENS, R. M. Sequence logos: a new way to display consensus sequences. *Nucleic acids research 18*, 20 (October 1990), 6097–6100.

[90] SCHONES, D. E., AND ZHAO, K. Genome-wide approaches to studying chromatin modifications. *Nature reviews. Genetics 9*, 3 (March 2008), 179–191.

[91] SEGAL, E., SHAPIRA, M., REGEV, A., PE'ER, D., BOTSTEIN, D., KOLLER, D., AND FRIEDMAN, N. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nature Genetics 34*, 2 (2003), 166–176.

[92] SHAO, J. An asymptotic theory for linear model selection. *Statistica Sinica 7* (1997), 221–264.

[93] SHEN-ORR, S. S., MILO, R., MANGAN, S., AND ALON, U. Network motifs in the transcriptional regulation network of escherichia coli. *Nature genetics 31*, 1 (May 2002), 64–68.

[94] SHI, P., AND TSAI, C.-L. Regression model selection: A residual likelihood approach. *Journal of the Royal Statistical Society. Series B (Statistical Methodology) 64*, 2 (2002), 237–252.

[95] SMYTH, G. K. *Limma: linear models for microarray data.* Springer, New York, 2005, pp. 397–420.

[96] SPELLMAN, P. T., SHERLOCK, G., ZHANG, M. Q., IYER, V. R., ANDERS, K., EISEN, M. B., BROWN, P. O., BOTSTEIN, D., AND FUTCHER, B. Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. *Mol Biol Cell 9*, 12 (Dec 1998), 3273–97.

[97] SPLETTSTOESSER, T. `http://en.wikipedia.org/wiki/File:Zinc_finger_DNA_complex.png`.

[98] SQUIDONIUS. `http://en.wikipedia.org/wiki/File:Microarray_exp_horizontal.svg`.

[99] STAM, F. J., MACGILLAVRY, H. D., ARMSTRONG, N. J., DE GUNST, M. C. M., ZHANG, Y., VAN KESTEREN, R. E., SMIT, A. B., AND VERHAAGEN, J. Identification of candidate transcriptional modulators involved in successful regeneration after nerve injury. *European Journal of Neuroscience 25*, 12 (2007), 3629–3637.

[100] STORMO, G. D. DNA binding sites: representation and discovery . *Bioinformatics 16*, 1 (2000), 16–23.

[101] STUART, J. M., SEGAL, E., KOLLER, D., AND KIM, S. K. A Gene-Coexpression Network for Global Discovery of Conserved Genetic Modules. *Science 302*, 5643 (2003), 249–255.

[102] SUGIURA, N. Further analysis of the data by akaike's information criterion and the finite corrections. *Communications in Statistics - Theory and Methods 7*, 1 (1978), 13–26.

[103] SZPARA, M., VRANIZAN, K., TAI, Y., GOODMAN, C., SPEED, T., AND NGAI, J. Analysis of gene expression during neurite outgrowth and regeneration. *BMC Neuroscience 8*, 1 (2007), 100.

[104] TEIXEIRA, M. C., MONTEIRO, P., JAIN, P., TENREIRO, S., FERNANDES, A. R., MIRA, N. P., ALENQUER, M., FREITAS, A. T., OLIVEIRA, A. L., AND SÁ-CORREIA, I. The yeastract database: a tool for the analysis of transcription regulatory associations in saccharomyces cerevisiae. *Nucleic Acids Res 34*, Database issue (January 2006).

[105] TIBSHIRANI, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B 58* (1994), 267–288.

[106] TROYANSKAYA, O., CANTOR, M., SHERLOCK, G., BROWN, P., HASTIE, T., TIBSHIRANI, R., BOTSTEIN, D., AND ALTMAN, R. B. Missing value estimation methods for dna microarrays. *Bioinformatics 17*, 6 (June 2001), 520–525.

[107] TSAI, H.-K., LU, H. H.-S., AND LI, W.-H. Statistical methods for identifying yeast cell cycle transcription factors. *Proceedings of the National Academy of Sciences of the United States of America 102*, 38 (2005), 13532–13537.

[108] TU, B., KUDLICKI, A., ROWICKA, M., AND MCKNIGHT, S. Logic of the yeast metabolic cycle: temporal compartmentalization of cellular processes. *Science 310*, 5751 (2005 Nov 18), 1152–8.

[109] VAN DER LAAN, M., AND RUBIN, D. Targeted maximum likelihood learning. *International Journal of Biostatistics 2*, 1 (2006), 1043–1043.

[110] VAN DER LAAN, M. J. Statistical inference for variable importance. *The International Journal of Biostatistics 2*, 1 (2006).

[111] WARD, L. D., AND BUSSEMAKER, H. J. Predicting functional transcription factor binding through alignment-free and affinity-based analysis of orthologous promoter sequences. *Bioinformatics 24*, 13 (2008), i165–171.

[112] WARNER, J. B., PHILIPPAKIS, A. A., JAEGER, S. A., HE, F. S., LIN, J., AND BULYK, M. L. Systematic identification of mammalian regulatory motifs' target genes and functions. *Nat Meth 5*, 4 (April 2008), 347–353.

[113] WASSERMAN, W. W., AND SANDELIN, A. Applied bioinformatics for the identification of regulatory elements. *Nat Rev Genet 5*, 4 (April 2004), 276–287.

[114] WERHLI, A. V., GRZEGORCZYK, M., AND HUSMEIER, D. Comparative evaluation of reverse engineering gene regulatory networks with relevance networks, graphical gaussian models and bayesian networks. *Bioinformatics 22*, 20 (Jul 2006), 2523–2531.

[115] WHITFIELD, M. L., SHERLOCK, G., SALDANHA, A. J., MURRAY, J. I., BALL, C. A., ALEXANDER, K. E., MATESE, J. C., PEROU, C. M., HURT, M. M., BROWN, P. O., AND BOTSTEIN, D. Identification of Genes Periodically Expressed in the Human Cell Cycle and Their Expression in Tumors. *Mol. Biol. Cell 13*, 6 (2002), 1977–2000.

[116] WHITTAKER, J. *Graphical Models in Applied Multivariate Statistics*. John Wiley & Sons Ltd., 1990.

[117] WU, W. S., AND LI, W. H. Identifying gene regulatory modules of heat shock response in yeast. *BMC Genomics 9*, 1 (2008).

[118] WU, W.-S., AND LI, W.-H. Systematic identification of yeast cell cycle transcription factors using multiple data sources. *BMC Bioinformatics 9*, 1 (2008), 522.

[119] XIAO, Y., AND SEGAL, M. R. Identification of yeast transcriptional regulation networks using multivariate random forests. *PLoS Comput Biol 5*, 6 (06 2009), e1000414.

[120] XIE, X., LU, J., KULBOKAS, E. J., GOLUB, T. R., MOOTHA, V., LINDBLAD-TOH, K., LANDER, E. S., AND KELLIS, M. Systematic discovery of regulatory motifs in human promoters and 3[prime] utrs by comparison of several mammals. *Nature aop*, current (February 2005).

[121] ZHANG, N. R., WILDERMUTH, M. C., AND SPEED, T. P. Transcription factor binding site prediction with multivariate gene expression data. *The Annals of Applied Statistics 2*, 1 (2008), 332–365.

[122] ZINZEN, R. P., GIRARDOT, C., GAGNEUR, J., BRAUN, M., AND FURLONG, E. E. M. Combinatorial binding predicts spatio-temporal cis-regulatory activity. *Nature 462*, 7269 (November 2009), 65–70.

[123] ZOU, H., HASTIE, T., AND TIBSHIRANI, R. On the "degrees of freedom" of the lasso. *Annals of Statistics 35*, 5 (2007), 2173–2192.